PREDICTION OF LIKELIHOOD OF FAILURE OF UNDERGROUND LINEAR ASSETS USING SURVIVAL ANALYSIS

By

SEE HYIIK TING

Bachelor of Science in Civil Engineering

University at Buffalo, The State University of New York

Buffalo, New York

2009

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2012

PREDICTION OF LIKELIHOOD OF FAILURE OF

UNDERGROUND LINEAR ASSETS USING

SURVIVAL ANALYSIS

Thesis Approved:

Dr. Hyung Seok Jeong

Thesis Adviser

Dr. Michael Phil Lewis

Dr. Rifat Bulut

Dr. Sheryl A. Tucker

Dean of the Graduate College

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER II

CHAPTER III

CHAPTER IV

Figure Page

CHAPTER I

INTRODUCTION

1.1 Background

America has been renowned for its great infrastructure since the post World War II era. A lot of the infrastructure systems were built then, and have significantly deteriorated by now. Some older systems have been in place for up to a hundred years, and the growing American population continues to place increasing demands on them (Powell, 2010). We can see roads and bridges getting older and developing cracks, but what about the water pipes below the ground? Imagine drinking water from such an old pipe.

Water infrastructure includes drinking water and wastewater infrastructure systems. In 2009, The American Society of Civil Engineers (ASCE) Report Card for America's Infrastructure gave these systems a D- grade. While utilities are facing the challenge of keeping up with deteriorating assets, they have to make sure that the Clean Water Act (CWA) requirements are being met (GHD, 2010). The CWA contains regulations and quality standards for water that is discharged into the waters of America. Polluted water is not only unsafe for drinking; it is a threat to many activities including fishing and swimming (USEPA, 2009).

America's water infrastructure system is in deep financial crisis. It has the highest projected shortfall of $108.6 billion, after roads and bridges that are estimated at $549.5 billion, for the next 5 years (ASCE, 2009). As we wait, the price tag will become more and more costly.

Utilities need an investment strategy that represents the best integration of maintenance, operations, and capital investment, where this integration delivers sustained performance at an acceptable level of service, at the lowest total cost of ownership, and at a level of risk the community is willing to tolerate. At the same time, utilities also have to ensure safety, security and resilience of the systems. This makes it necessary to include risk management into asset management practices (Santora and Wilson, 2008).

1.2 Problem Statement

Risk management helps an asset manager to decide the best time to intervene in an asset's lifecycle. That refers to maintenance, rehabilitation, or replacement activities. The timing of these activities greatly impacts the effectiveness of these assets. In some cases, it is best to take action before an asset fails. In other cases, it is otherwise. Hence, asset managers have questions such as:

- Which assets, and how likely will they fail this year, in 5 years, in 10 years, etc.
- Should current operations and maintenance activities be improved or should an asset be renewed now
- Should investments be proactive (take action before an asset fails) or reactive (take action after an asset fails)
- When and how much should be invested in inspections and condition assessments
- How can changes in risks be accounted for

Risk is the product of likelihood of failure (LoF) and consequence of failure (CoF). In light of realizing the need for risk management, many methods have been used to determine the LoF of an asset, which is the dominant component of risk. Thus, the determination of LoF is critical as it greatly affects management decisions. Current common practices include using decision tree and table based methods. A decision tree looks like a map of branches that displays the procedure to determine LoF. It leads the asset manager through a series of decisions and concludes with a subjective level of risk. The table based method is a table where the asset manager uses qualitative assessment – very high, high, medium, low, and very low, to assess the LoF and CoF and decides the level of risk.

There is a lot of subjectivity in both decision tree and table based methods. This can lead to a lot of variation in the level of risk depending on how the asset manager interprets each situation at hand. Thus, there can be a lot of problems and bias in justifying what actions to take. Therefore, a better way of determining LoF is needed for more effective determination of risk and asset management. Specifically, a more science based method is required to predict LoF. Among some quantitative methods, survival analysis has many advantages. It is able to use various factors to predict the LoF and it accounts for incomplete information, such as that faced by underground assets. These assets are seldom tracked and certain information is difficult to obtain due to their nature of being buried. Survival analysis can also account for data that change over time. Therefore, it directly addresses the needs of utilities.

In short, current methods to determine underground linear asset LoF are too subjective. Asset managers need to justify their decisions about what to do when, and to which assets. A more advanced method is needed to help determine underground linear asset LoF.

1.3 Objective

The objective of this research is to develop survival models to help asset managers better determine the LoF of underground water and sewer pipes. Water data will be collected from the Greater Cincinnati Water Works (GCWW) while sewer data will be collected from the Metropolitan Sewer District of Greater Cincinnati (MSD).

1.4 Tasks

The following tasks have to be completed to accomplish this research's objective.

1. Review current methods used to assess asset risk and determine asset LoF.
2. Develop parametric and non-parametric survival models for water and sewer pipes to predict their LoF or survival probability at a given age.
3. Determine the best parametric model for both water and sewer pipes comparing Weibull, Exponential, Lognormal, Gamma and Loglogistic distributions.
4. Compare the performance between parametric and non-parametric models.
5. Compare the performance between models developed for assets that are carefully grouped and assets that are not grouped.

CHAPTER II

PREVIOUS STUDIES RELATED TO DETERMINING END OF ASSET LIFE

This chapter reviews previous studies that have been done in the area of determining LoF of an underground linear asset. It is a major part of determining risk and helps an asset manager decide when to intervene in an asset's lifecycle. The methods used can be simply grouped into judgment based and data based methods.

2.1 Judgment Based Methods

Judgment based methods are, as its name implies, based on judgment or experience. If relevant historical data are not available, the asset manager should consider these non-statistical techniques and/or embark upon a program to capture the relevant data. Non-statistical methods may use expert judgment and subjective assessment with structured evaluation criteria. One such technique is the table-based method, which uses the percentage of asset physical life that has been consumed to estimate remaining life.

The first step in this process is to evaluate the percentage of physical life used in terms of 5 performance aspects:

- Technical performance (structural condition)
- Operational performance
- Reliability
- Availability
- Maintainability

These aspects of performance are rated according to the following table:

Table 2.1: Performance Aspects (GHD)

Element	**Description**					
SCORE	**1**	**3**	**5**	**7**	**9**	**10**
Technical Performance	Substantially exceeds current requirements	Exceeds current requirements	Meets current requirements but with room for improvement	Obvious concerns: cost/benefit questions	Inefficient; becoming ineffective, obsolete	Failing, not capable of sustaining required performance
Operational Performance	Negligible attention required	Exceeds current requirements	Meets current requirements but with room for improvement	Obvious concerns: cost/benefit questions	Difficult to sustain performance	Failing, not capable of sustaining required performance
Reliability	As specified by manufacturer	Infrequent breakdown	Occasional breakdown	Periodic breakdown	Continuous recurrent breakdown	Virtually inoperable
Availability	Virtually always operational	Out of service only for very short periods	Out of service for moderate period; moderately difficult to return to service	Increasingly difficult to return to service; parts becoming a challenge	Extensive downtime duration; difficult to return to service; parts, difficult to acquire, rare skills required	Virtually impossible to return to service; parts no longer available; unavailable trained personnel
Maintainability	Easily maintained; OEM maintenance is straightforward	Largely preventive maintenance with some corrective maintenance beginning to show up; baseline monitoring	Increasing minor maintenance required; periodic corrective maintenance including some repair shortening of monitoring intervals	Scheduled maintenance becoming frequent; more experienced trades people required for maintenance; frequency of work orders increasing substantially with short monitoring intervals	Work orders well above average for type of asset; recurrent minor repair; close monitoring required; most senior people required to sustain performance	Maintenance is frequent with recurrent patterns of failure; asset must be virtually constantly monitored to sustain performance
% Physical life consumed	Almost new; up to 10% physical life consumed	Up to 30% physical life consumed	Up to 50% physical life consumed	Up to 70% physical life consumed	Up to 90% physical life consumed	Virtually consumed, imminent failure
SCORE	**1**	**3**	**5**	**7**	**9**	**10**

The performance scores of all 5 aspects are averaged and converted to percent effective life consumed according to the following table:

Table 2.2: Converting Performance Scores to % Life Consumed (GHD)

% Effective Life Consumed	10%	20%	30%	40%	50%	60%	70%	80%	90%	Failed
Composite Asset Performance Score	1	2	3	4	5	6	7	8	9	10

This percentage of effective life consumed is then converted into LoF in a 1 to 10 scale according to the following table.

Table 2.3: Converting % Life Consumed to LoF (GHD)

% Effective Life Consumed	10%	20%	30%	40%	50%	60%	70%	80%	90%	Failed
LoF	1	2	3	4	5	6	7	8	9	10

For example, if the average performance score of an asset is 4.5, the effective life consumed would be 45% and the LoF would be 4.5 on a scale of 10 (GHD).

This method is good when data is very limited or not available. However, judgments can vary among users and this can produce very different results.

2.2 Data Based Methods

Data based methods use data to prove an outcome. These methods require a substantial amount of historical data to make a prediction. Data based models are model driven, where an attempt is made to find the best model that explains data that is analyzed. It usually involves a probability

distribution and model parameters that explain the relationship between input and output data. Models will then be calibrated using maximum likelihood methods and goodness of fit tests will be done to validate them. Data based methods include statistical models, and soft computing models which will be discussed (Marlow, Davis, Trans, Beale and Burn, 2009).

2.2.1 Statistical Methods

Eight statistical methods will be reviewed here. They are failure event data-based, service lifetime-based, cohort-survival, ordinal regression, Markov Chain, Bayesian, deterministic, and physical probabilistic.

The first one is failure event data approaches. These models are applied where recorded failure data is available. They predict failure rates of asset groups. The asset groups are formed using information such as pipe length, pipe diameter, installation date, breakage history, soil type, etc. Shamir and Howard developed an exponential equation considering only pipe age (one factor) to predict the number of breaks of water pipes. They first classified assets by material, construction method, soil, temperature, and pressure conditions, and suggested that an equation can be developed for each homogenous group. They combined these prediction values with cost data to determine the best time to repair or replace assets. They concluded that this methodology can help in decision making but not replace good judgment (Shamir and Howard, 1979).

The next type is service lifetime approaches. These models are very similar to failure event data approaches but are used with service lifetime data. They are also applied to asset groups. The difference between these models and failure event data models is that these models only consider the time to first failure. Herz developed a probability model and predicted renewal or service rates

for different water pipe groups in Europe. Similarly, the groups had to be homogenous or containing pipes in similar environments. It was reported that maintenance and repair activities can extend the life of pipes (Herz, 1996).

Next are survival models. These models predict the probability of survival within a condition over a certain number of years. They require condition data and predict the number of years it might take for an asset to transition into a worse state. This information helps with knowing when an asset will be in critical condition and with planning inspection activities. These models are applied to asset groups. Two most popular distribution assumptions used in survival models are the exponential and Weibull distributions. In Duchesne et al.'s work, the parametric method of survival analysis has been done for various asset group sizes. They have found that the exponential model is more suitable for smaller sample sizes (1000 samples or less) compared to the Weibull model (Duchesne, Beardsell, Villeneuve, Toumbou and Bouchard, 2012). In Syachrani's paper, the non-parametric method of survival analysis has been used on underground pipes, eliminating the need of any distribution assumption. He had also found that the non-parametric method was not suitable for equipment failures (Syachrani, 2010). In this paper, a comparison between the parametric and non-parametric methods will be included.

Next are ordinal regression methods. They are similar to survival models and require condition data and asset attributes. They are also applied to asset groups. These models assume that deterioration is a continuous process and threshold values are used to distinguish assets with different conditions. Several researchers have used these models to classify assets into good or bad conditions. Davies et al. investigated the factors affecting sewer condition in a London utility using logistic regression. They found that sewer section length, size, location, material, depth, use, background soil properties, local ground water regime, and traffic flow were significant factors. However, they found that root penetration, burst history of adjacent water pipes, and age were insignificant, which surprised them (Davies, Clarke, Whiter, Cunningham and Leidi, 2001).

Ariaratnam et al. studied sewer pipes from Edmonton, Canada and found that age, diameter and waste type were significant variables. Their model was able to predict the probability of a sewer system being in a deficient state (Ariaratnam, El-Assaly and Yang, 2001). Both these research acknowledged that collecting enough quality data was a challenge.

The next type is Markov Chain approaches. They also use condition data. The uniqueness about Markov Chain models is their assumption that deterioration occurs as steps from one condition to another. The probability of transitioning from one state to a worse one depends on age and operating environment. These models can be applied to both individual assets and asset groups. A Markov chain model predicts the probability that pipes in one condition state progress to other condition states, as defined by a transition matrix. It assumes that the progress is discrete, not continuous, and that all the assets grouped together will have the same transition matrix. The Markov Chain theory recognizes that the next state of an asset depends only on its current state, not its previous. Likewise, a future condition depends on its current condition. If an asset is currently at condition state 1 out of 5, it is more likely to progress into state 2 than states 3, 4, or 5. This likelihood to change from one state to another is known as the Markovian transition probability. In a Markovian deterioration model, the transition probabilities and deterioration rates can be estimated by experts using condition assessment data. Estimating transition probabilities is the biggest challenge in using the Markov model. Baik et al. proposed the ordered probit model to estimate transition probabilities in a Markov deterioration model. Sewer data from the City of San Diego was used. The results did not perform very well in the goodness of fit tests. The team suggested that they lacked quality data. In order to improve their results, they needed continuous condition assessment data (Baik, Jeong and Abraham, 2006).

Next are Bayesian approaches. These models forecast the probabilities of failure or condition states of asset groups. The Bayes' Theorem provides a relationship between the probability that an initial prediction is correct after the addition of new data and the previous estimates before the

new data is added. It allows previous estimates of the asset conditions to be combined with inspected data that is available so that new predictions can be made about the conditions of the asset group. The challenge of these models is in grouping the assets using suitable criteria. Kulkarni et al. learned that assets need to be grouped into various condition states. Their team studied deterioration of gas pipelines and developed a model combined with cost data. The result was the Cast Iron Maintenance Optimization System for the gas industry that can help asset managers make decisions on managing their pipeline assets (Kulkarni, Golabi and Chuang, 1986).

Deterministic approaches are used when the relationship between input variables and failure rates are known. Deterministic models are divided into empirical and physical approaches.

The empirical approach uses equations to fit a set of data. One example is regression analysis. Shamir and Howard developed a regression equation as follows to predict failure rate of water pipes:

$$N(t) = N(t_0)\exp[A(t+g)] \tag{2.1}$$

where t is the time in years from present; $N(t)$ is the failure rate per unit length per year; $N(t_0)$ is the failure rate at year of installation; g is the age of the pipe at time t and A is the coefficient of failure rate (in year^{-1}) (Shamir and Howard, 1979). Empirical deterministic models must be applied to homogenous groups of assets that have historical failure data. Asset related data such as pipe length, pipe diameter, and installation date are also needed to group the assets. While Shamir and Howard developed a simple exponential model using one variable to predict water pipe failure rates, Clark et al. improved the model. They showed that the time to first failure follows a linear pattern but once a pipe starts requiring maintenance, its maintenance rate increases exponentially (Clark, Stafford and Goodrich, 1982). Some disadvantages of these

models are that fitting the equations to observations can be very challenging, they may only be applied to homogenous groups of assets, and they cannot account for time-dependent factors.

Physical deterministic models are mechanism-based models that explain degradation and failure processes. Therefore, a critical requirement is information on asset deterioration. For example, these models have been used to predict time to corrosion failure of water pipes. Variables that describe the assets' degradation have to be obtained, such as the corrosion rate. Physical deterministic models usually predict the service lifetime (rather than failure rate) of an individual asset. Randall-Smith et al. (1992) developed a model to predict corrosion failure of water pipes:

$$\rho = \left(\frac{t}{P_e + P_i}\delta\right) - t \tag{2.2}$$

where ρ is the remaining service life; t is the age of the pipe; P_e is the external pit depth; P_i is the internal pit depth; and δ is the thickness of the original pipe wall. One problem with this model is that a linear assumption was used for corrosion rate (Marlow, Davis, Trans, Beale and Burn, 2009).

Physical probabilistic models can be used when historical failure data are not available. These models use small samples to study actual deterioration and degradation. They are based on load-capacity relationships that study loading conditions and their effects on failure such as corrosion. These models require information on pipe material, operating loads (internal and external), and condition data. The uncertainties within variables are represented using probability distributions. This can be done using techniques such as Monte Carlo simulation. Then, the results can be applied to a network of assets. Examples of uncertainties are variations in soil electrochemistry, water chemistry in contact with a pipe, and defects in pipe processing (Marlow, Davis, Trans, Beale and Burn, 2009).

Monte Carlo simulation is used to estimate probability functions together with an underlying physical model. In a Monte Carlo simulation, random variable values are generated continuously. Each value is then used to predict a failure time. This is done until a certain number of trials is reached or until the standard error of the mean predicted lifetime is below a certain designated value. Davis et al. used this simulation to estimate the probability of longitudinal fracture in Asbestos Cement pipes(Davis, De Silva, Marlow, Moglia, Gould and Burn, 2008).. The predicted lifetimes are then fitted to a probability distribution, which is finally used to estimate failure probability over time. One of the most popular probability distributions used is the Weibull distribution, due to its two-parameter property that makes it fit datasets well. The challenge in using this approach is in gathering actual asset deterioration and degradation data. However, it is attractive because it can be used on newer assets. Statistical models that use historical data are more suitable for older assets.

2.2.2 Soft Computing Methods

Soft computing methods are data driven models. These models process inputted information in several steps or "layers" and create a connection with predicted output results. They use a lot of historical data for calibration and are tested using independent data.

Artificial Neural Network (ANN) is a type of soft computing approach that predicts values for individual assets and asset groups. It uses all variables that are thought to influence failure rates, such as pipe diameter, pipe length, pipe age, installation year and geographical location. Model coefficients are adjusted so that predicted outputs are as close to historical outputs as possible. These models are considered "Black Box" solutions and the computations behind the models are often unknown. Extra care should be taken when using these models because it is difficult to

judge how much data is ideal to make good predictions. Tran et al. used ANN to predict deterioration in stormwater pipes in Australia. They had approximately 650 data points and validated the model using Bayesian weight estimation and conventional back-propagation weight estimation. The prediction accuracies turned out to be only 69% and 58% respectively (Tran, Ng and Perera, 2007)

Next are fuzzy models. These models incorporate "vagueness" to compensate for processes that are difficult to understand, such as the deterioration process. For example, when pipes are given condition scores (e.g. 1-5 score), it is difficult to justify the boundaries of each score. Fuzzy models are usually used to "fuzzify" input data and are used with other models. Kleiner et al. (2006) have used this model. The table below shows an example of deterioration rate chart for given pipe ages and conditions:

Table 2.4: Fuzzy Rule Base for Deterioration Model (Kleiner, Sadiq and Rajani, 2006)

Pipe Condition (C):		**excellent**	**good**	**adequate**	**fair**	**poor**	**bad**	**failed**
Age (A):	new	slow	average	fast	very fast	very fast	very fast	very fast
	young	slow	average	fast	fast	fast	very fast	very fast
	medium	very slow	slow	average	average	fast	fast	very fast
	old	very slow	very slow	slow	slow	average	average	fast
	very old	very slow	very slow	very slow	slow	slow	average	average

For example, if a pipe is new and the condition is adequate, its deterioration rate is fast. However, if the same pipe is in fair condition, its deterioration rate is very fast. Every rate (very slow, slow, average, fast, and very fast) has an associated numerical value. By using such associations, different deterioration curves can be estimated for every asset throughout their lifecycle. In short, this method uses fuzzy logic and expert opinions to determine the LoF, rather than using data from a database. Figure 2.1 below illustrates the deterioration of a sewer pipe from age 20 to 40. The asset condition changes from good to adequate, fair, and poor. The lower half of the figure shows the condition rating at age 40. However, if another figure was provided for the condition

rating at age 20, the reader could tell that the area enclosed shifted from a good state to a worse state.

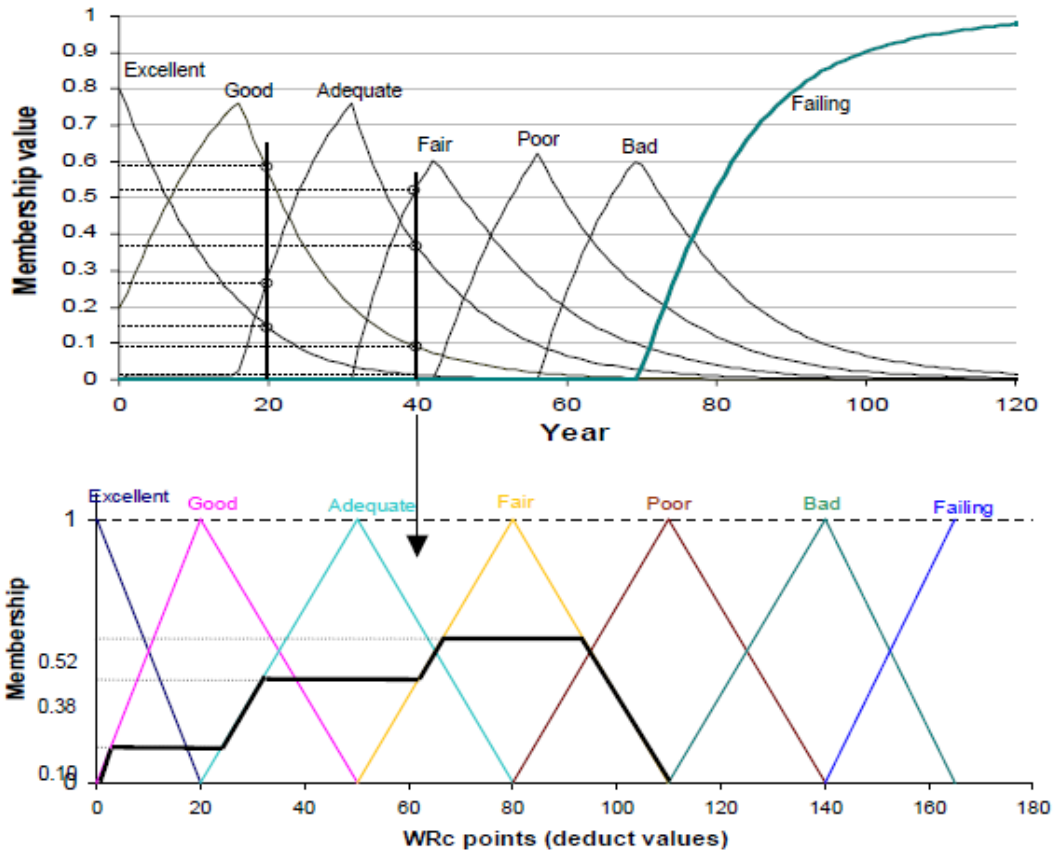


Figure 2.1: Fuzzy Model (Kleiner, Sadiq and Rajani, 2006)

One problem with these models is that they must be validated using assets that have been consistently inspected (Kleiner, Sadiq and Rajani, 2006).

CHAPTER III

PROCEDURE TO PREDICT LOF USING SURVIVAL ANALYSIS

This chapter presents the methodology and theoretical background involved to predict asset LoF. The use of this methodology requires an intermediate level of knowledge in statistics. The procedure is illustrated in a flowchart that can be found in Appendix B. The procedure involves two major stages; a) classification of assets by common behavior into "management strategy groups" and b) development of survival curves.

3.1 Data Collection, Quality Assurance, and Management

In this step, data is collected and checked for missing information, inconsistency, redundancy, etc. Data points are then recoded, replaced, or transformed whenever necessary. In other words, the master spreadsheet database is prepared so that a statistical analysis could be applied.

Data collection, quality assurance and management of data is the process of putting the data in order and getting rid of irrelevant data. This section explains what information needs to be collected, where to find them, and how to prepare them for analysis.

.

In order to predict the LoF, the asset manager must first have a clear definition of failure of each type of asset. Without a clear definition, consistency in the determination of failure and the specific date of failure cannot be readily determined.

There are three types of asset failures otherwise known as "end-of-asset-life" upon which an investment strategy has to consider:

Table 3.1: End-of-Asset Life Descriptions

End-of-Asset-Life	**Description**
End of physical life	An asset physically stops working, collapses, or is otherwise non-functional or non-operational.
End of service level / capacity life	An asset can no longer do what customers or operators require it to do (including reaching capacity limits).
End of economic life	An asset ceases to be the lowest cost alternative to satisfy a specified level of performance or service at an acceptable level of risk.

This research addresses physical failure. For example, many utilities now use the Pipeline Assessment Certification Program (PACP) scores when assessing sewer pipe conditions and plan their management strategies using the PACP assessment results. The PACP report is used to calculate the structural condition of sewer pipes using a five point scale 1, 2, 3, 4, and 5. Table 3.1 defines these scores.

Table 3.2: PACP Score Definitions (Elizabeth Ehret, 2011)

PACP score	**Definition**
1	Excellent; minor defects
2	Good; has not begun to deteriorate
3	Fair; moderate
4	Poor; will become grade 5 in near future
5	Immediate attention needed

Although immediate attention is needed for a grade 5 sewer pipe, it does not really mean that the pipe has collapsed. The pipe may still be functioning but based on PACP definitions, it is at critical condition and requires immediate attention due to defects such as cracking and sagging. Unless actual sewer pipe failure records are available, this grade 5 is considered as a failure in this study in order to apply statistical theories to determine the LoF. If the asset manager has a clear definition of failure and keeps good failure records (specifically, the systematic and consistent recordation of the date of failure), those data must be used to develop the LoF instead of the PACP based condition assessment data set.

Physical failure may be defined very differently for different assets. Different failures require different approaches. For pumps, if there is no pumping of fluid, it could be time to replace certain components such as the motor or fuse. However, if there is high current draw, rough cycle, or low pressure, maintenance activities such as cleaning or lubricating may be needed. Take a light bulb as another illustration. The bulb could fail due to a burn out, flashing, or dimming. It may require replacement or reconnection. Water pipes may fail due to leaks, low pressure, or corrosion while sewer pipes may fail due to leaks, corrosion, collapse, etc. It is very important that failure be clearly defined, because it affects operations and maintenance activities that are required.

In order to use a statistical method to estimate the LoF, the asset manager also needs to check whether sufficient amount of asset failure related data is available since any statistical method is a data intensive approach. Rich historical asset data may lead to accurate and reliable models in predicting the LoF. The recommended data for water and sewer pipes to determine physical failure include;

a) Physical attributes such as age, length, material, diameter, etc.
b) Performance data such as condition assessment data

c) Environmental attributes such as soil type, density of proximate trees, corrosivity, climate and freeze/thaw properties, groundwater level, etc. and
d) Operational attributes such as maintenance, repair, installation and failure records

Other available information might be an advantage. When only limited data are available, the statistical analysis may not be able to produce the most accurate and reliable result. However, limited data does not indicate that a statistical analysis cannot be carried out so long as interpretation of the results is within the constraints of the data.

For water and sewer pipes, physical attributes, installation and failure records or condition assessment data are the minimally required data for a basic level of statistical analysis. However, since survival analysis requires censoring, some data points with missing failure records or condition assessment data are required. For an advanced level of statistical analysis, additional data attributes such as environmental and other operational attributes are needed. Having a broader range of data attributes means that more factors can be considered to determine the LoF. Although some input data may not be useful, these insignificant input variables will be screened out by a statistical analysis procedure and only statistically meaningful input variables will remain and be used to develop the prediction model.

In collecting asset data, various databases such as Geographical Information System (GIS), Computerized Maintenance Management System (CMMS), condition assessment data, and others should be used. These data may be managed by different departments in a utility organization. So, coordinating efforts between departments are required. Then, these data have to be put together into a master spreadsheet based database. Various asset attributes from different datasets must be associated with each asset ID for further statistical analysis in developing a master database and this master database should be easily accessible. Figure 3.1 shows a sample master database with

the first few data points and some of its available attributes. It contains more than 50,000 sewer pipe entries with 23 different attributes arranged nicely in one spreadsheet.

	A	B	C	D	E	F	G	H	I	J
1	No.	UniqueID	Diameter	Material	Length	Slope	Age	Condition	StreamX	HighwayMajR
2	1	0332115:0332114	8	VCP	279.246	3.79	43	5	0	0
3	2	0425099:0425098	15	PVC	190	0.86	16	1	1	0
4	3	0105038:0105035	8	VCP	272.005	0.52	14	2	1	0
5	4	0403045:0403044	8	VCP	98.078	31	30	3	0	0
6	5	0142045:0142043	8	VCP	339.386	4.4	49	5	0	0
7	6	0102065:0102064	8	VCP	74.507	0.14	55	5	0	0

Figure 3.1: Master Database

After collecting data, they need to be prepared for analysis. For example, data for "material" might need to be recoded so that specific numbers can represent different types of materials. This is because statistical software recognizes numbers better than words. As another example, "Length" data may need to be checked to ensure the units are consistent. Obvious typographical errors in the database such as "2.00" instead of "200" or instances when alphabets were in place by mistake should also be corrected.

It is also important to recognize that a pipe may have multiple failure records. These records can be treated independently if the failures do not occur on the same spot. In most cases, only the first failures are of interest in an analysis, because that shows how long a new pipe can last before more resources have to be invested.

Figure 3.2 shows how failure records can be treated independently.

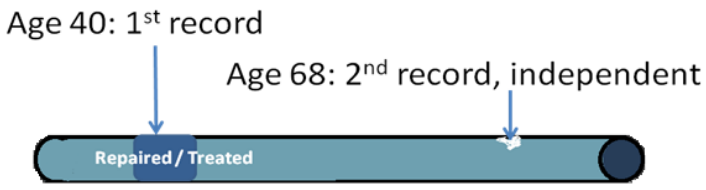


Figure 3.2: Treatment of Failure Records

On a particular pipe, the 1st failure is recorded at age 40. After 28 years, a 2nd failure is recorded on the same pipe. Although these 2 failures occurred on the same pipe, they can be used as 2 independent data points if the asset manager decides so – one failure at age 40 and another at age 68.

3.2 Asset Classification

The next step is to classify assets into Management Strategy Groups (MSGs) for effective management in the long run. The following explains how MSGs are created.

MSGs tend to be sub-groups of asset classes. Each MSG is expected to contain assets that are likely to display a similar behavior over their lifetime, especially with respect to aging and failure patterns. These similar behavior patterns are a composite expression of three separate groups of factors that drive failure:

- Core or intrinsic factors
    - Steady, continuous deterioration mechanics related to engineering design and specification, material, manufacturing processes, manufacturer, construction/installation management

- Operating Environment factors
    - Operating condition mechanisms such as temperature, corrosivity, nature of material being handled, weather exposure, soil moisture, soil chemistry, depth, soil type and aggressiveness, proximity to electrical fields, etc.
- Operational factors
    - O&M factors such as quality, nature and frequency of maintenance, nature and timing of renewal, historic rate of failure

For example, consider an inventory of gravity pipes that contains reinforced concrete pipes (RCP). The utility knows that RCP gravity pipes that were installed before 1950 and that tend to have high H_2S content have a maximum potential life of 75 years. Then, the following criteria may be used to group these pipes as a unique cluster (this grouping effectively divides the asset class "RCP gravity pipe" into two behavior groups – one defined as follows and one for all else.

[Type] = Gravity Pipes

[Material] = Contains RCP

[Install Year] = <1950

[H_2S Category] = True

When MSG group classification fields like these are determined, MSGs can be set, and models can be developed for each group of assets. The models should be able to explain the MSGs well. In other words, if a model is developed to predict the LoF of a certain MSG, the model should reasonably predict the LoF of each asset in that MSG. Otherwise, either the model or the MSG criteria needs to be redeveloped for more effective management.

MSGs facilitate the development of management strategies that are fine tuned for the behaviors observed. The utility might have existing MSGs that could be used. However, the MSGs'

designations could be obsolete or have not been effective. In such cases, the utility might want to validate or redevelop MSGs to be used for analysis.

There are generally two methods to develop MSGs; 1) Judgment based MSGs and 2) Statistics Based MSGs. Judgment based MSGs use the management team's professional judgment, experience and engineering knowledge to classify assets. It should be a deliberative process where the team most knowledgeable about the asset gathers and discusses criteria to be used to classify assets. When statistics based MSGs are developed, the existing asset related data needs to be fed into a statistical program to indicate groups of assets or clusters.

Using a judgment based method is beneficial because the team's expertise and years of experience can be clearly reflected into classifying asset groups. However, the results of judgment based MSGs cannot be validated without empirical verification. The statistics based method is certainly more evidence-based (what we think we know and what the data really show are not always the same) and could be used by a less experienced team. However, it will likely require at least a modest level of statistical knowledge and experience and access to a statistical package. Interpretation of the results could also be a challenge for users. However, this method can easily be validated since statistics is based on numerical facts. In other words, validly rendered statistics based results are truly 'correct' since there are supporting numerical facts to identify the groups.

3.2.1 Judgment Based Classification

In judgment based classification, a series of discussions within the utility involving asset managers, maintenance engineers, operation engineers and field superintendents may be necessary and are encouraged to capture their judgment and experience in developing the criteria.

It is important to note that a high number of criteria will lead to a large number of MSGs and may reduce the number of assets in each MSG but will improve the accuracy and reliability of subsequent prediction models such as the LoF for each MSG. However, too many criteria may lead to inefficient management of MSGs simply because there are too many MSGs to manage. A small number of criteria will increase the number of assets in each MSG and increase the variability of deterioration and failure patterns, leading to lower accuracy and lower reliability of prediction models. However, a small number of MSGs will be easy and simple for the utility to manage. Thus, a reasonable number of criteria should be used as a trade-off value between the efficiency of management and the accuracy/reliability of prediction models.

Once the MSG criteria are established, the utility can classify various assets into different groups according to the criteria. All assets have to belong to a certain MSG. External studies may also be conducted to find MSG criteria that can be used. For example, an asset manager may find that practice or research performed by other utilities or practitioner groups show certain criteria to be relevant in helping to classify his assets.

Table 3.2 shows a sample of judgment based MSGs developed for a real utility in the U.S. This utility used material type, diameter, installation year, and soil corrosivity as the criteria for developing MSGs for water pipes. Note that specific years were used to classify CIPs and DIPs. This is due to major specification changes that occurred to these pipes in certain years and the utility is confident that this change would affect the deterioration and failure patterns of the pipes. It is also important to note that this type of judgment may only be obtainable from experienced engineers, field personnel and managers. Due to a lack of historical data, a judgment based

approach is appropriate. A statistics based approach, which will be explained in the next section, cannot detect these important changes unless they are recorded as a data attribute in the database.

Table 3.3: Sample of Judgment Based MSGs

Material Type	**Diameter and Installation Year**	**Soil Corrosivity**
Cast Iron Pipe (CIP)	CIP-large diameter (≥ 15")	High / Medium /Low
	CIP, small diameter (< 15"), pre 1955	High / Medium /Low
	CIP, small diameter (< 15"), 1955-1966	High / Medium /Low
	CIP, small diameter (< 15"), post 1966	High / Medium /Low
Ductile Iron Pipe (DIP)	DIP_P1_(pre 2002)	N/A
	DIP_P1_(post 2002)	N/A
Pre-stressed Concrete Cylinder Pipe (PCCP)	PCCP all	N/A

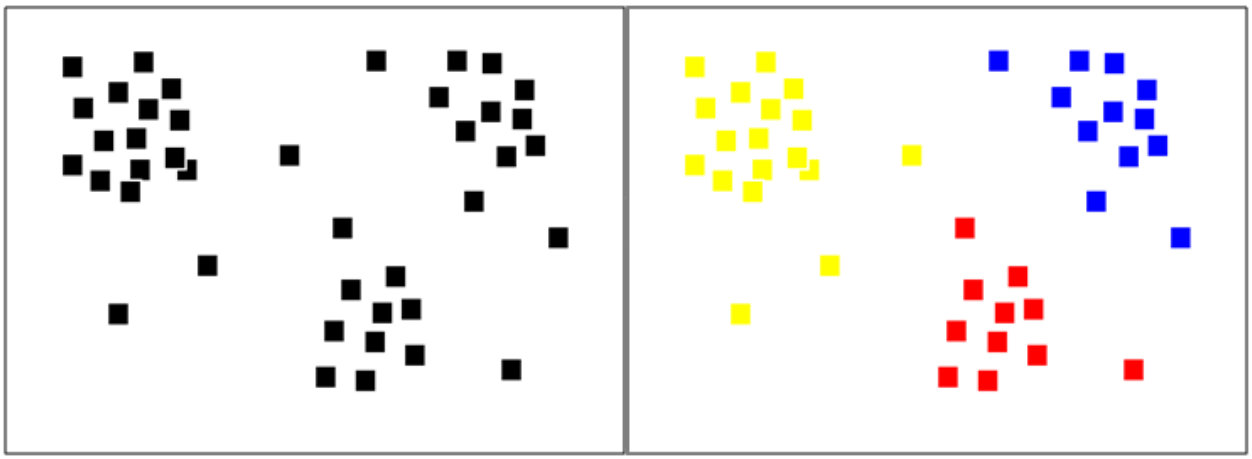
3.2.2 Statistics Based Classification

There are several ways to classify data statistically. In general, a statistics based approach could be either data driven or model driven. Data driven methods find natural groupings in the data, while model driven methods attempt to separate data into predefined groups based on a combination of practice based knowledge (the body of engineering science, for example) and emerging science. To have predefined groups is equivalent to using the judgment based method. Therefore data driven methods are discussed here since the judgment based method has already been explained in the previous section.

Statistical clustering techniques are advanced data driven methods that create virtual groups of assets with a maximum degree of association among each other using statistical software and

numerical computation. This method should likely be more “accurate” than the judgment based method, but is not necessarily more efficient.

Figure 3.3 shows an example of how statistical clusters may look like. It can be seen that there are three clusters in the data. Statistical software can calculate the distances between data points and, using sophisticated mathematics, identify clusters in the dataset. As shown in Figure 3.3b, three clusters have been identified using different colors. However, without systematic analysis, clusters are not evident, as shown in Figure 3.3a.



(a) (b)

Figure 3.3: Statistical Clustering

There are several different clustering methods, such as hierarchical clustering, k-means clustering and factor analysis. The difference among these methods is the computational concepts used to identify the clusters. Ultimately, the clustering process groups data that are similar in some fashion with each other.

Table 3.3 shows a real sample of MSGs developed for a utility in the U.S. through a statistical clustering method. In total, six clusters were identified for each type of sewer pipe material. It shows that sewer pipes under highway crossing and right of way (ROW) may have different aging and failure patterns than pipes buried under a large number of trees. Also, pipes under

single family residential area may have different deterioration patterns than pipes under commercial and restaurant zones. This type of grouping of assets may not be apparent if only a judgment based classification method is used.

Table 3.4: Sample of MSGs Developed Using Clustering Method

Clusters	**Frequency count (%)**	**Strong characteristics**
Cluster A	33.88%	Single family residential area
Cluster B	29.83%	Non-Single family residential area
Cluster C	21.05%	Trees
Cluster D	8.34%	Multifamily residential area
Cluster E	5.08%	Commercial, restaurant
Cluster F	1.83%	Highway crossing, transportation, ROW
Total	100.00%	

After the classifications have been confirmed and MSGs are created, survival curves may be developed. Any necessary modifications or adjustments to the MSGs should be performed at this step. No changes should be made to the MSGs subsequently during the analysis process.

3.3 Survival Model Development

In this step, the survival model is developed for each MSG. The process involves selecting significant variables, assuming distributions, and developing survival curves.

Failure often has many different potential causes (the progression of mechanical events that leads to a typical failure is called the “failure mode”; an asset can have many different failure modes). The challenge to the asset manager is to identify those failure modes that account for most

failures for a given class of assets in a designated operating environment. Recalling the three separate groups of factors that drive failure is helpful here to organize which variables ("causes") to pursue:

- Core or intrinsic factors
    - Steady, continuous deterioration mechanics related to engineering design and specification, material, manufacturing processes, manufacturer, construction/installation management
- Operating Environment factors
    - Operating condition mechanisms such as temperature, corrosivity, nature of material being handled, weather exposure, soil moisture, soil chemistry, depth, soil type and aggressiveness, proximity to electrical fields, etc.
- Operational factors
    - O&M factors such as quality, nature and frequency of maintenance, nature and timing of renewal, historic rate of failure

The above list is generic; not all of the listed variables are significant in producing the bulk of failures. Which variables are significant vary from agency to agency and even zone to zone or plant to plant within an agency. Some variables may be correlated with other variables. For example, a larger sewer pipe will obviously have a lower velocity compared to a smaller pipe that carries the same amount of sewer in the same network. In such a scenario, only one of the two variables is needed. By selecting more significant variables, the efficiency of the analysis process may be improved. In addition, insignificant variables only contribute a very small percentage to the outcome as compared with significant variables. Therefore, only the variables that are significant should be selected and used.

Figure 3.4 here illustrates the concept of significant variables.

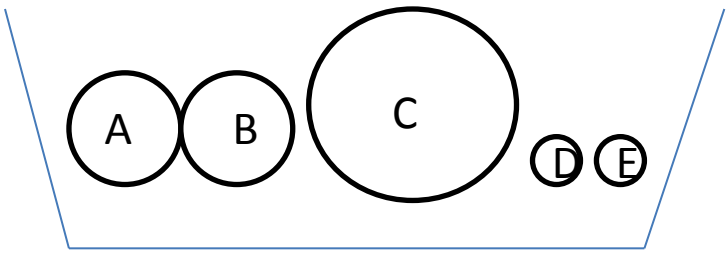


Figure 3.4: Concept of Significant Variables

Assume that all five circles A, B, C, D, and E are in a basket and they represent different weights. Larger circles weigh more than smaller circles. We want to estimate the weight that the basket is carrying, but our weighing scale cannot accommodate the whole basket with all the weights at once. Therefore, the weights have to be measured one at a time. After measuring A, B and C, the added weights yield 15 pounds. Weight D is 0.12 pounds and weight E is 0.11 pounds. Finally, we can conclude that the basket is carrying 15 pounds, because D and E are insignificant to some degree. Certainly, this process helps us eliminate possible insignificant variables to improve the efficiency of the analysis.

There are many methods that can be used to select significant variables. Among them are stepwise regression, forward selection, and backward elimination. This process has to be done for each MSG that has been determined. To verify the selected variables, be sure that all their p-values are small (e.g. <0.05), implying significance. A p-value of 0.05 implies a 95% confidence level, or that the chance of being incorrect is 5%. Similarly, a p-value of 0.1 implies a 90% confidence that the variable is significant. A maximum acceptable p-value should be defined.

Figure 3.5 shows a sample result of the initial process of selecting significant variables using the backward elimination method for a selected utility. There were 15 variables, in which all of them were tested simultaneously:

- DIAMETER: Pipe diameter (in.)

- LENGTH: Pipe length (ft.)
- SLOPE: Pipe slope
- STREAMX: Pipe is located under stream crossing, 1 is yes
- HIGHWAYM: Pipe is located under highway or major road crossing, 1 is yes
- RAILROAD: Pipe is located under rail road crossing, 1 is yes
- LAKEWETL: Pipe is located under lake or wetland, 1 is yes
- RESTAURA: Number of surrounding restaurants
- ROOTPROB: The presence of root problem, 5 is worst
- SLUDGEPR: The presence of sludge problem, 5 is worst
- DEBRIPRO: The presence of debris problem, 5 is worst
- GREASEPR: The presence of grease problem, 5 is worst
- JOINTPRO: The presence of joint problem, 1 is yes
- COLLAPSE: Pipe with collapsed section, 1 is yes
- BROKENPR: Pipe with broken section, 1 is yes

From the initial test, it was found that variable RAILROAD has a p value of 0.9998, which is the most insignificant variable. To refine the result, variable RAILROAD can first be removed before running the analysis again. Then, variables “LAKEWETL”, “JOINTPRO”, “BROKENPR”, and “DIAMETER” were also removed since they have no data points. The process continues where the most insignificant variable is removed every time the analysis is repeated, until all remaining variables are significant.

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.3300	0.2161	3.9064	4.7536	401.38	<.0001
LENGTH	1	-0.0005	0.0004	-0.0014	0.0003	1.46	0.2275
SLOPE	1	-0.0545	0.0322	-0.1176	0.0087	2.86	0.0908
STREAMX	1	-0.2924	0.1288	-0.5450	-0.0399	5.15	0.0232
HIGHWAYM	1	0.2522	0.2108	-0.1609	0.6653	1.43	0.2314
RAILROAD	1	10.7279	41027.08	-80400.9	80422.32	0.00	0.9998
LAKEWETL	0	0.0000	.	.	.	.	.
RESTAURA	1	-0.6418	0.2963	-1.2225	-0.0612	4.69	0.0303
ROOTPROB	1	0.0830	0.0640	-0.0425	0.2084	1.68	0.1950
SLUDGEPR	1	-0.0685	0.0673	-0.2005	0.0634	1.04	0.3088
DEBRIPRO	1	-0.0374	0.1261	-0.2846	0.2099	0.09	0.7669
GREASEPR	1	0.1326	0.1798	-0.2199	0.4851	0.54	0.4609
JOINTPRO	0	0.0000	.	.	.	.	.
COLLAPSE	1	-0.8448	0.5293	-1.8821	0.1925	2.55	0.1104
BROKENPR	0	0.0000	.	.	.	.	.
DIAMETER	0	0.0000	.	.	.	.	.
Scale	1	0.4938	0.0455	0.4122	0.5915		
Weibull Shape	1	2.0253	0.1866	1.6906	2.4262		

Figure 3.5: Selecting Significant Variables

In this example, the final variables remaining that are significant are STREAMX and RESTAURA with p-values of 0.0323 and 0.0169 respectively as shown in the figure below.

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.0725	0.0662	3.9428	4.2022	3786.03	<.0001
STREAMX	1	-0.2852	0.1332	-0.5462	-0.0241	4.59	0.0323
RESTAURA	1	-0.7338	0.3073	-1.3361	-0.1316	5.70	0.0169
Scale	1	0.5182	0.0477	0.4327	0.6206		
Weibull Shape	1	1.9297	0.1775	1.6114	2.3109		

Figure 3.6: Selecting Significant Variables

Forward selection method however starts with no variables in the model. As variables are being added, their degrees of significance are being checked. The process stops when an added variable causes one or more of the variables to be insignificant. The model must end with all significant variables.

Stepwise regression is a combination of both forward selection and backward elimination and is the most popular method. It starts with no variables in the model. As variables are being added, the resulting model is checked to see if any variable included gives significant contribution and if the contribution is independent of other variables (that is, it is not measuring the same thing as another variable already in the equation). In other words, the variable(s) in the model have to be significant and independent. If the contribution from a particular variable is very small compared to those of other variables or is insignificant, that variable should be dropped.

In most statistical software packages, the process of selecting significant variables is relatively simple. Therefore, comparisons can be done between the results from different methods easily.

After significant variables are selected, a distribution assumption has to be made for each MSG for the development of parametric models.

The essence of survival analysis is its associated distribution (curve) of "deaths" – in our case the failure distribution curve. Survival data may take on many distribution patterns. It may, for example, take the shape of the common "bell" curve with a peak (center) and symmetrical "sides", or it may have one side dominating the other ("skewed"). The challenge is to capture the shape of the distribution curve in a mathematical statement. Survival analysis is a set of mathematical theories used to do this. Doing so systematically allows for the accurate projection of the "how many, when" question raised before (in section 1.2).

Survival analysis has been widely used in reliability studies. Its aim is to predict the probability that an asset can continue to function as intended for a given period of time, given its current conditions. There are two types of survival models – parametric and non-parametric. The parametric model has specific assumptions about the distribution of survival data. When the right assumptions are used, the results are more accurate. However, the non-parametric models do not rely on any distribution assumptions. These models will be further explained throughout this paper.

For the parametric model, a group discussion among team members would be an excellent way to come up with a general consensus regarding any typical failure pattern of a specific MSG. The LoF is related to the survival function. So, if the utility feels confident that the MSG failure pattern follows a certain distribution (e.g. Exponential, Weibull, and Lognormal) based on expert judgments, then a parametric method can be used to develop a survival model for predicting the LoF. Below are some samples of Exponential, Weibull and Lognormal distributions:

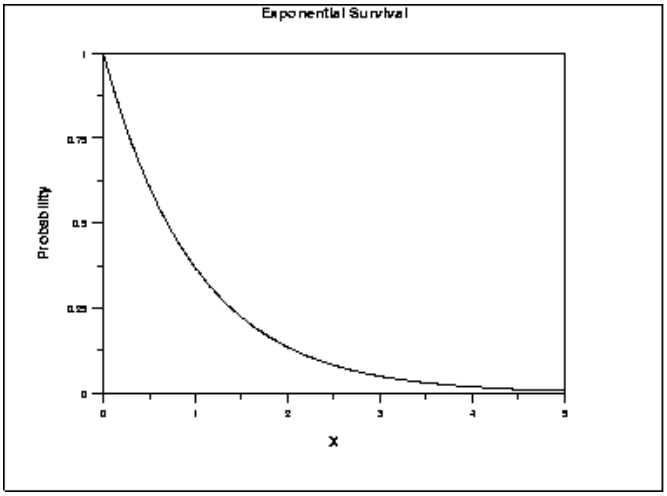


Figure 3.7: Exponential Distribution (NIST, 2012)

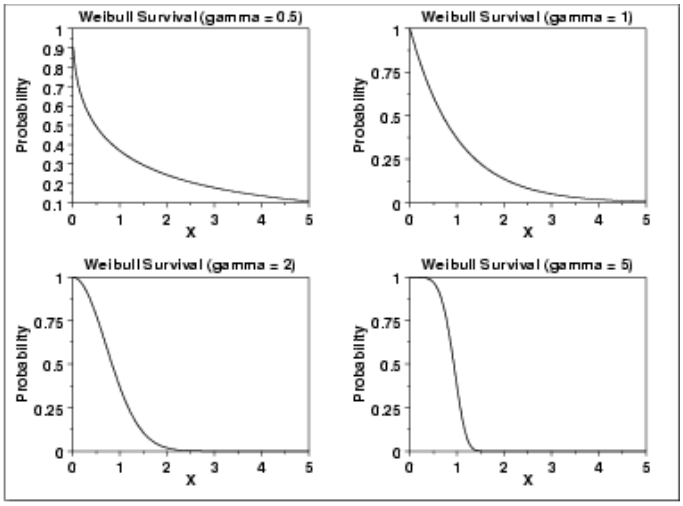


Figure 3.8: Weibull Distribution (NIST, 2012)

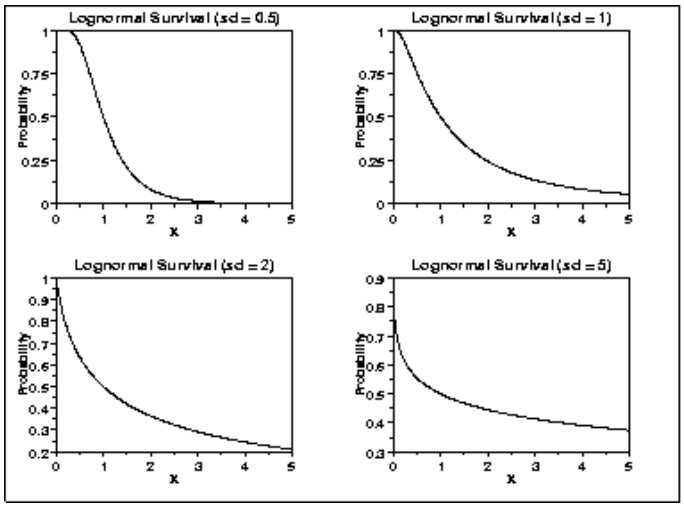


Figure 3.9: Lognormal Distribution (NIST, 2012)

The parametric method has several advantages over the non-parametric method.

- Standard errors (inaccuracy) of parameter estimates tend to be smaller
- Closed-form (predefined) expressions of the hazard and survival functions are available
- Good results can be obtained even in the case of small sample sizes
- Estimation procedure is more efficient if the assumed distribution model is correct

However, in the case of non-parametric models, no assumptions of the underlying distribution are required. This method uses the natural distribution it finds from the given data. These models are more flexible and are able to deal with any probability distribution. The disadvantage of the non-parametric approach is that it needs much more data to get reasonable results. Although there is no certain threshold value, it is understood that if this method is chosen, the results depend on how much data is available. The more data available, the more accurate results would be. It is also more difficult to get estimates of functions that are needed for this method.

When using statistical software to perform analysis, the procedural differences between parametric and non-parametric methods are minimal, however, except for the results. Since the non-parametric method finds the natural distribution in the data, it may be more accurate.

However, the parametric method is mainly attractive because it typically yields smaller standard errors and is much more efficient.

Next, the survival curves are to be generated.

In using survival analysis, several functions are involved:

- Hazard Function
- Survival Function
- Probability Density Function

The hazard function is also known as a risk or mortality rate. It is the instantaneous rate at which an event such as failure happens. For example, if an asset records 1 failure in 10 years, its hazard rate would be 1/10 or 0.1, assuming constant hazard over those 10 years. Of course, the hazard may not be constant for such a long period of time for most assets. As a comparative statement, when we say that a car is traveling at 70 miles per hour it does not mean that the car will travel exactly 70 miles after one hour. This statement is valid only if the speed of the car is kept constant. The hazard function can be used to derive survival or probability density function, and vice versa. The hazard functions are different for parametric and non-parametric models.

The survival function gives the probability that an asset survives past a certain time t. The plot it generates is the survival curve. The survival function is also called reliability function and is related to the hazard function such that if the failure rate is increasing, the probability of survival past time t will decrease. It is important to note that the survival function is a probability while the hazard function is a rate. As an example, the survival function can also tell how many failures would have occurred by 60 years. The general equation for survival function is defined by:

$$S(t) = \exp[-\int_0^x h(u)du], \tag{3.1}$$

where $h(u)$ is the hazard function.

The survival function can also be defined as:

$$S(t) = \exp(-\exp(-\mu\alpha)\, t^{\alpha}) \tag{3.2}$$

where $\mu = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \sigma\varepsilon_i$. Here, $\beta_0, \ldots, \beta_k$ are estimated parameters. $x_{i1}, \ldots, x_{ik}$ are the values of k covariates. $\varepsilon_i$is a random error term, and σ is the scale parameter, and α is $1/\sigma$. The survival function tells the probability of survival for any pipe of age *t.*

The probability density function in survival analysis is the probability of event occurrence, or failure, at various times t. The popular normal distribution which is a bell-curve is an example of probability density function. The probabilities (of occurrence) of each value (on the x-axis) increase to a peak and then decrease again. At the peak, the probability is at its highest (0.5), and, consequently, the corresponding value is known as the expected value.

The following explains how the related functions are developed. Parametric survival analysis is similar to ordinary linear regression. Assuming T_i to be a random variable representing survival time for the *i*th individual in the sample, and let $x_{i1}, \ldots, x_{ik}$ be the values of k covariates for that same individual, the model is then:

$$logT_i = \beta_o + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \sigma\varepsilon_i \,, \tag{3.3}$$

or

$$T_i = \exp(\beta_o + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \sigma\varepsilon_i) \,, \tag{3.4}$$

where $\varepsilon_i$is a random error term, and $\beta_0, \ldots, \beta_k$and σ are parameters to be estimated.

When an assumption can be made about the distribution of survival time for any particular MSG, parametric survival analysis will be performed. Here is a list of parametric models, and every distribution has different hazard and survival functions:

- Exponential
- Weibull
- Extreme Value
- Lognormal
- Gamma
- Birnbaum-Saunders

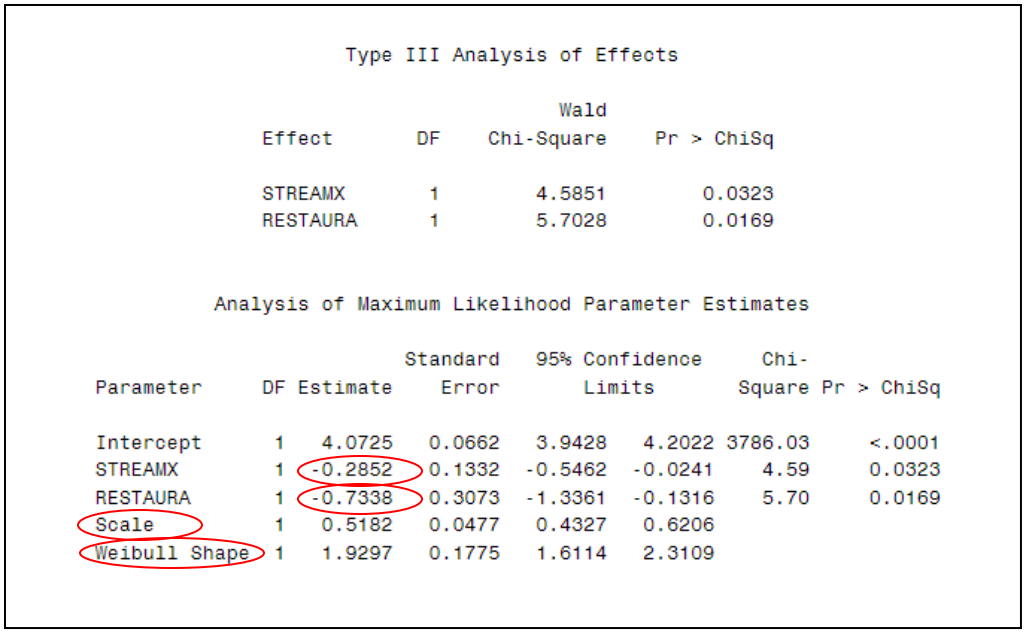
However, if the non-parametric method is chosen, tools such as the Cox proportional hazards model or Kaplan-Meier estimator is used to develop hazard and survival functions.

Based on the selected parametric or non-parametric method, the associated functions should be developed. These functions contain several parameters such as the shape, scale, shift, and others depending on the model. To estimate these parameters, several methods may be used:

- Graphical Estimation
- Maximum Likelihood
- Method of Moments
- Least Squares
- Probability Plot Correlation Coefficient (PPCC) and Probability Plots

The most popular methods are the graphical estimation and maximum likelihood. When using statistical software, these estimation procedures are not explicitly revealed. The statistical software program will perform the calculations in the background and show only the results.

Figure 3.10 shows the results of a parametric analysis using the Weibull model. The two significant variables selected and used here are STREAMX and RESTAURA. Notice that besides estimates for the two variables and the intercept, there are also estimates for the Weibull scale and shape factors.



Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
STREAMX	1	4.5851	0.0323
RESTAURA	1	5.7028	0.0169

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.0725	0.0662	3.9428	4.2022	3786.03	<.0001
STREAMX	1	-0.2852	0.1332	-0.5462	-0.0241	4.59	0.0323
RESTAURA	1	-0.7338	0.3073	-1.3361	-0.1316	5.70	0.0169
Scale	1	0.5182	0.0477	0.4327	0.6206		
Weibull Shape	1	1.9297	0.1775	1.6114	2.3109		

Figure 3.10: Parametric Analysis Using Weibull Distribution

It can be deduced that for every increment in STREAMX (stream crossing) there will be a 24.8% decrease in survival time. The calculation is as follows: 100[exp(-0.2852)-1] = -24.8%

Similarly, for every increment in RESTAURA (surrounding restaurant present), there will be a 52% decrease in survival time. This also implies that this variable is more significant than STREAMX because it affects the survival time by 52% as opposed to 24.8%. The calculation is as follows: 100[exp(-0.7338)-1] = -52%

The scale estimate is σ. Changes in this estimate may affect the shape (compress / expand) of the hazard function, depending on the type of distribution. Since it is between 0 and 1 in this Weibull model, it implies that the hazard is increasing and survival time is decreasing.

The Weibull Shape parameter is the reciprocal of the scale parameter. It has no special use; some statisticians prefer it over the scale estimate.

After these parameters have been estimated, reliability data should be plotted. There are a few types of plots:

- Hazard and Cumulative Hazard plots
- Survival plots
- Cumulative Distribution Function (CDF) plots

The figures in the next few pages show examples of hazard plot, cumulative hazard plot, and survival plot and the ways to interpret them. Figure 3.11 shows the hazard plot for the Weibull distribution. The plot is smooth, and the hazard increases at a constant rate. At point A, which is about age 30, the hazard or failure rate is 0.02 / year, or 1 in 50 years. At point B, which is about age 60, the failure rate increases to 1 in 26 years. This is the failure rate of assets in this MSG over time. Operations and maintenance activities can be planned by defining the maximum allowable failure rate.

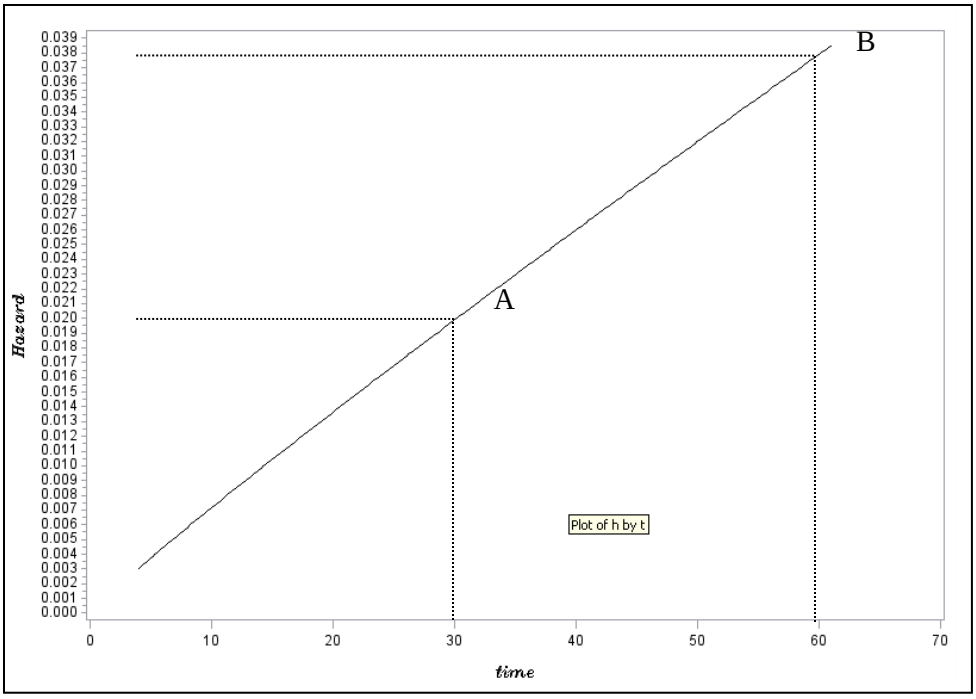
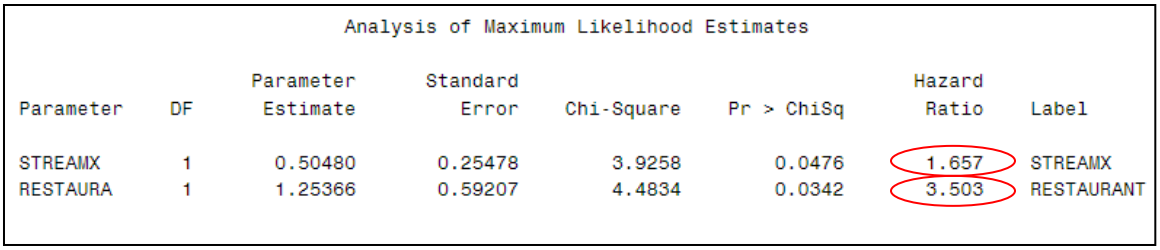


Figure 3.11: Hazard Plot for Weibull Distribution

The figure below shows the results of a non-parametric analysis using Cox Regression model. Notice that the analysis results show hazard ratios for each variable.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
STREAMX	1	0.50480	0.25478	3.9258	0.0476	1.657	STREAMX
RESTAURA	1	1.25366	0.59207	4.4834	0.0342	3.503	RESTAURANT

Figure 3.12: Non-Parametric Analysis Using Cox Regression

It can be inferred that for each increase in STREAMX there is a 65.7% increase in hazard. The calculation is as follows: (1.657-1) x 100 = 65.7%.

Similarly, for each increase in RESTAURA there is a 250.3% increase in hazard. The calculation is as follows: (3.503-1) x 100 = 250.3%.

Figures 3.13 and 3.14 show sample plots of the cumulative hazard and survival function from the non-parametric analysis above. It can be seen that this MSG's survival probability drops rapidly from age 18 to 40. Therefore, the asset manager would probably plan for more inspection or maintenance activities before this age range. Between ages 40 and 57, the survival probability is quite stable. If the utility's goal is to perform major rehabilitation to assets that fall below the survival probability of 0.4, it can definitely be deduced that this activity can be done the latest at age 57. There is no need to perform major rehabilitation at age 40 because the survival probability is quite stable up to age 58.

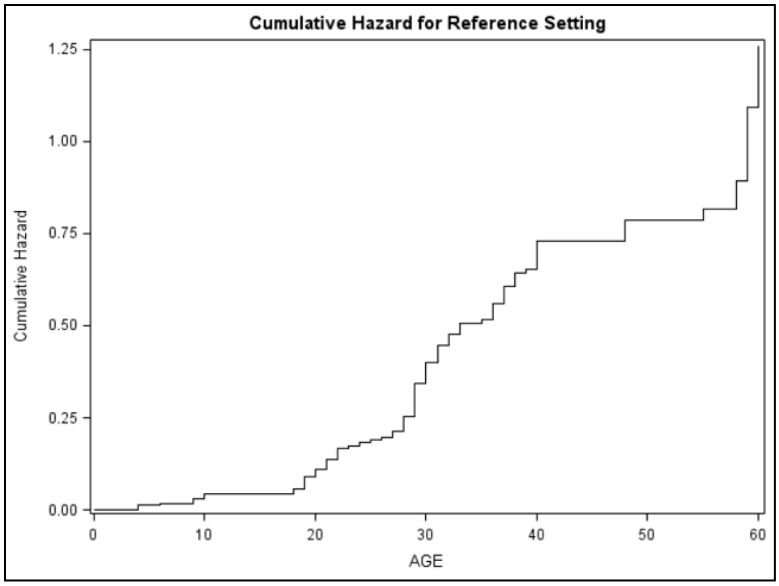


Figure 3.13: Non-Parametric Cumulative Hazard Plot

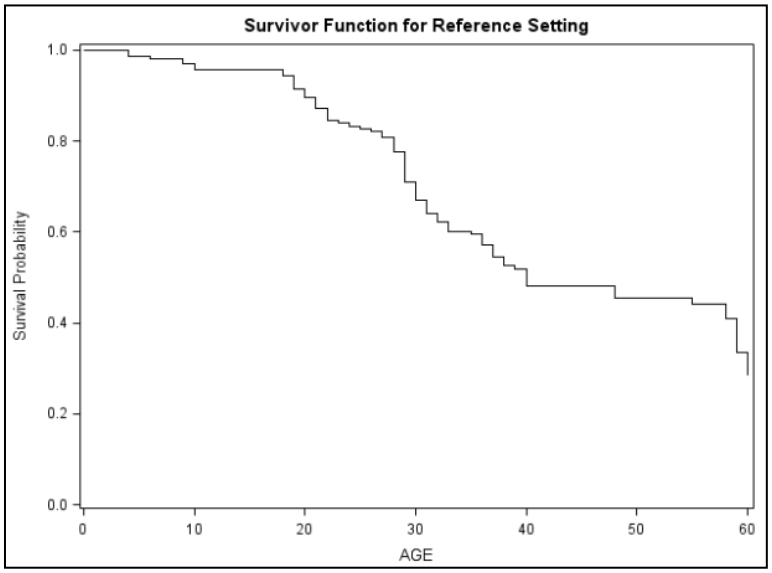


Figure 3.14: Non-Parametric Survival Plot

The cumulative hazard plot shows that the failure rate of the asset group is increasing at an increasing rate from age 18 to 40, since it is curved upward. The failure rate increases at a decreasing rate from age 40 to 57.

3.4 Model Validation

The figure below shows a distribution curve and some data points. The basic concept of model validation is to measure the distance between the data points and the curve. The smaller the distances, the better a curve fits.

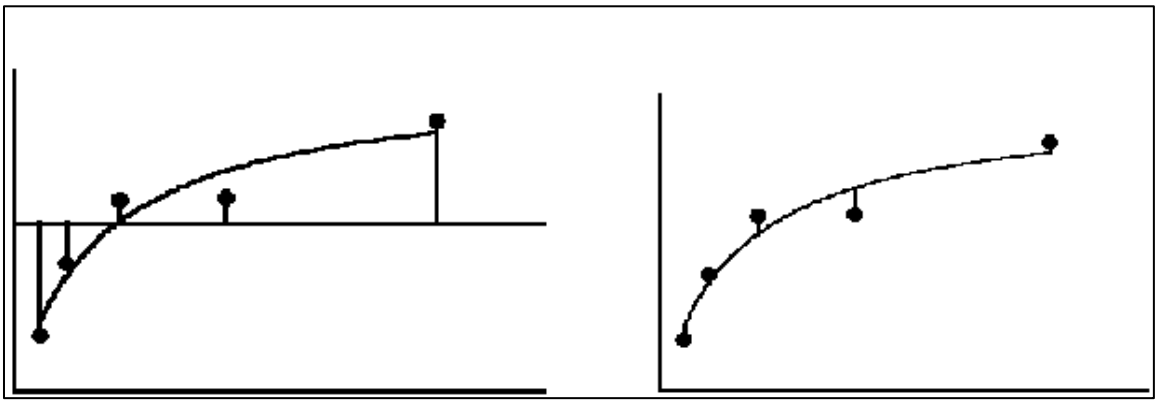


Figure 3.15: Distribution Curve

Numerical or graphical methods can be used to validate a survival model. Numerical methods use fit statistics such as "-2Loglikelihood" and "AIC". They are usually used to compare several models. Lower statistics mean a better fit.

Graphical methods are plots that show how well a model fits. Figure 3.16 shows a graphical method of the previous example parametric analysis, called the probability plot. If all the data points are on the straight line, the assumed distribution is perfect. If they are within the shaded area, the fit is within the 95% confidence limit.

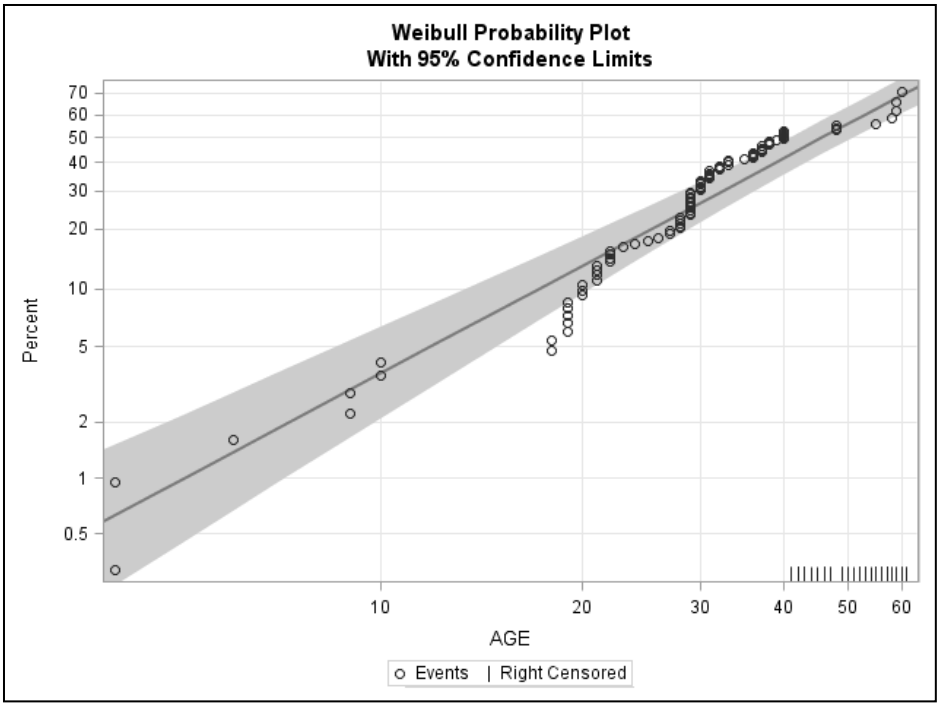


Figure 3.16: Probability Plot for Weibull Distribution

It can be seen that a Weibull distribution assumption may not be the best (less than 95% confidence) for this example since there are data points lying outside the shaded area. We might have to consider a different distribution assumption.

Figure 3.17 shows a repeated analysis using a Gamma distribution. This distribution fits better and should be chosen over Weibull distribution, because all the data points fall within the shaded area.

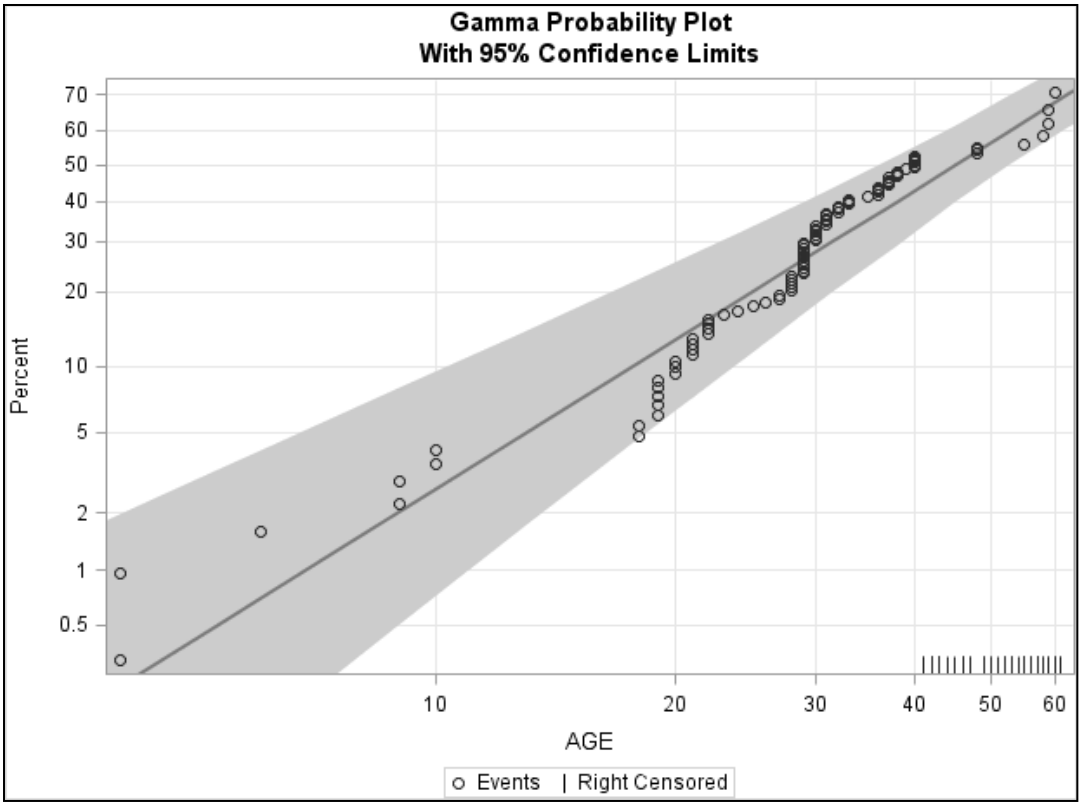


Figure 3.17: Probability Plot for Gamma Distribution

If the data points do not fit the plot nicely, either new MSGs have to be defined or the model has to be redeveloped. Otherwise, the validated model is ready to be applied to asset networks.

3.5 Model Application

A validated model is ready to be applied to the entire network of assets in the same MSG. This model should be used together when reviewing CoF score and risk mitigation strategies to help the asset manager decide whether the risk exposure is within tolerable limits or not. It will also help estimate when this limit will be exceeded, and hence what short and long term decisions should be made.

CHAPTER IV

RESULTS OF PREDICTING LOF USING SURVIVAL ANALYSIS

This chapter presents the results of survival analysis for predicting end of asset life for both water and sewer pipes.

4.1 Water Pipe Data

The water pipe data was provided by the Greater Cincinnati Water Works (GCWW). This section presents the analysis results of water pipe data. Failure was defined by any repair or rehab activities that were done on a pipe. These water pipes were interfered due to leaks, pressure losses, or other unusual symptoms were discovered either from regular inspections or from their users. In this analysis, only the first failures on each pipe were used.

4.1.1 Data Collection, Quality Assurance, and Management

This section describes the data attributes that were collected from GCWW data. Table 2 shows the data attributes of the water data that was collected from GCWW.

Table 4.1: Major Data Attributes of Collected Water Data

Attribute	**Attribute Type**	**Definition**
Asset ID	Physical	5 or 6 digit asset identifier unique for each main segment (for example, 74967, 107433)
Serviceare	Other	Area the pipe is located at (for example, EH)
Lifecycle	Other	Identifies a main segment as active or abandoned
Strlabel	Other	Street name
Installyea	Operational	Installation year of pipe segment
Report Date	Operational	Date of reported failure
Material	Physical	Material of pipe (AC, CI, CU, DI, etc.)
Diameter	Physical	Diameter of pipe in inches
C_factor	Other	Calculated Hazen Williams coefficient
Pressure	Other	Operating pressure of pipe segment
Elevation	Other	Estimated elevation of pipe from GIS contours
Shape_leng	Physical	Length of pipe in feet calculated by GIS
Deadend	Other	Identifies main segment as plugged or capped
Pitodist	Other	Pitometer district used to manage leak surveys
100Scale	Other	Identifies specific 100-scale record drawing of portion of distribution system
Groundsurf	Environmental	Ground surface type above pipe segment
Administra	Other	Administrative area in which pipe segment resides
Fixed_asse	Physical	Number assigned to all main segments related to a specific capital project
Neighborho	Other	Neighborhood where pipe is located
Size_wm	Physical	Diameter of pipe in inches
Joint_type	Physical	Type of joint (flange, compression, lead, etc.)
Break_desc	Other	Description of failure (corrosion, circular crack,

		longitudinal crack, etc.)
Out_pipe_c	Other	Exterior description of pipe condition
In_pipe_c	Other	Interior description of pipe condition
Outside_co	Other	Internal number representing Out_pipe_c information
Inside_con	Other	Internal number representing In_pipe_c information
Main_break	Other	Internal number representing Break_desc
Address	Other	Address of pipe location
Recno	Other	Internal record number for maintenance activity
Year_recno	Other	Year recorded for maintenance activity

The water data had 30 variables. Those that had duplicated information, missing information, irrelevant data or were the same for every asset were disregarded. In the end, 10 variables were used in the analysis. They were “Asset ID”, “Installyea”, “Report Date”, “Material”, “Diameter”, “C_factor”, “Pressure”, “Elevation”, “Shape_leng”, and “Neighborho”.

4.1.2 Asset Classification

GCWW’s master spreadsheet had over 17,000 data points. In this study, a simple judgment based classification was being done where pipe material and size were assumed to be good criteria. After discarding data points with missing installation year records, there were 9,886 remaining. There were 7 different types of materials, of which the majority were cast iron. The material distribution is shown in Figure 4.1.

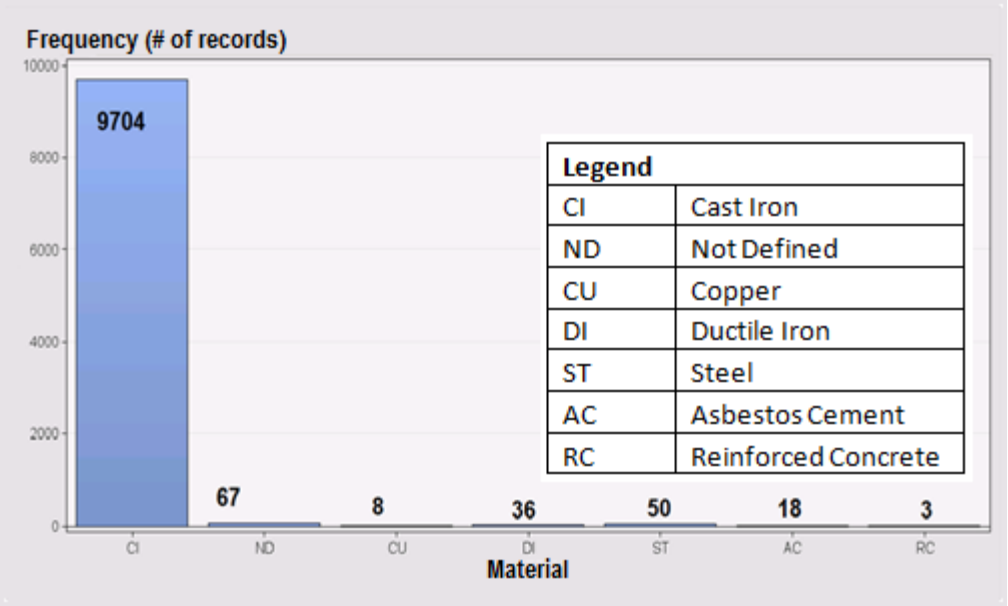


Figure 4.1: Material Distribution of GCWW Records

The cast iron pipes were selected and it was found that the majority were 6 inches in diameter. Therefore, this group of records was selected for further analysis. Figure 4.2 shows the diameter distribution of cast iron records, after further analysis and removal of duplicate records. The 6 inch cast iron records were still the majority, by far.

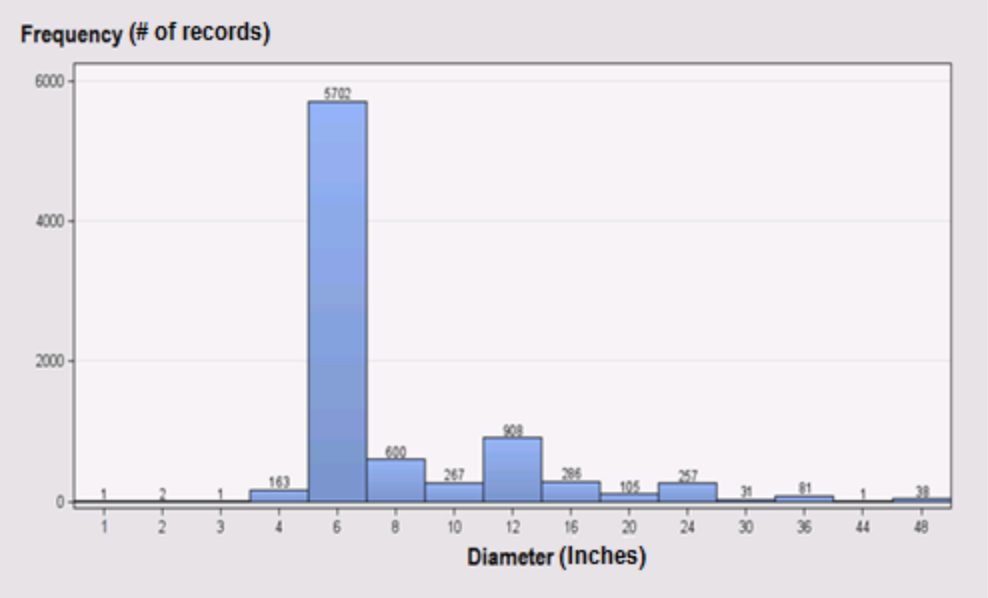


Figure 4.2: Diameter Distribution of Cast Iron Pipe Records

The MSG confirmed for water pipes were 6" cast iron pipes. The number of good data points was 5,702, from 5,245 pipes. This was because some pipes had failed more than once.

A subset of these water pipes was also selected to be another MSG. The figure below shows the distribution of data points in different neighborhoods. Most of the data points did not have neighborhood information, except for 971 of them. There were altogether 53 different neighborhoods. Out of these data points, 60 of them were from the Bond Hill neighborhood, which were the majority. Although this was not a very big number, it had been selected to test the effect of grouping. Only the first failures in this dataset were used. In other words, failure records of pipes that failed more than once were disregarded. There were 48 data points for Bond Hill neighborhood being used. The "Other" slice of the pie chart shows the total number of data points from all other minority neighborhoods.

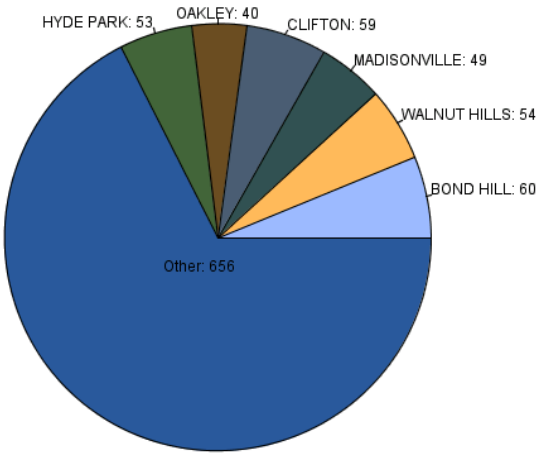


Figure 4.3: Neighborhood Distribution of Water Pipes

4.1.3 Survival Model Development

This section shows the results of developing parametric and non-parametric survival models for the 2 groups of water pipes –5,702 cast iron pipes and 48 Bond Hill Neighborhood pipes.

4.1.3.1 Parametric Survival Model

The 6” cast iron water pipes had 4 variables that were potentially affecting physical survival time– “C_FACTOR”, “PRESSURE”, “ELEVATION”, and “SHAPE_LENG”. The results of selecting significant variables are shown in Figure 4.4.

Before

Analysis of Maximum Likelihood Parameter Estimates						
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-3677.28	974.8899	-5588.03 -1766.53	14.23	0.0002
C_FACTOR	1	-0.0050	0.0005	-0.0060 -0.0041	105.07	<.0001
Pressure	1	8.9520	2.3700	4.3070 13.5971	14.27	0.0002
ELEVATION	1	3.8763	1.0262	1.8650 5.8876	14.27	0.0002
Shape_Leng	1	-0.0007	0.0000	-0.0007 -0.0006	294.84	<.0001
Scale	1	0.2314	0.0061	0.2198 0.2436		
Weibull Shape	1	4.3212	0.1132	4.1049 4.5489		

After

Analysis of Maximum Likelihood Parameter Estimates						
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-3677.28	974.8899	-5588.03 -1766.53	14.23	0.0002
C_FACTOR	1	-0.0050	0.0005	-0.0060 -0.0041	105.07	<.0001
Pressure	1	8.9520	2.3700	4.3070 13.5971	14.27	0.0002
ELEVATION	1	3.8763	1.0262	1.8650 5.8876	14.27	0.0002
Shape_Leng	1	-0.0007	0.0000	-0.0007 -0.0006	294.84	<.0001
Scale	1	0.2314	0.0061	0.2198 0.2436		
Weibull Shape	1	4.3212	0.1132	4.1049 4.5489		

Figure 4.4: Selection of Significant Variables for Water Pipes

Before any selection was done, all the variables were tested. It can be seen that all the 4 variables were tested to be significant. Therefore, these 4 variables – “C_FACTOR”, “PRESSURE”, “ELEVATION”, and “SHAPE_LENG” were selected for this analysis. Here, we can deduce that the survival time, S can be expressed as S = exp [-3677.28 – 0.005(C_Factor) + 8.952(Pressure) + 3.8763(Elevation) – 0.0007(Shape_Leng)]. In the statistical software used, the survival curve was then generated to show survival probabilities.

Figure 4.5 shows the parametric survival curve for 6" cast iron water pipes. Here, the Weibull distribution was first assumed because of its flexibility to fit datasets, and its popularity.

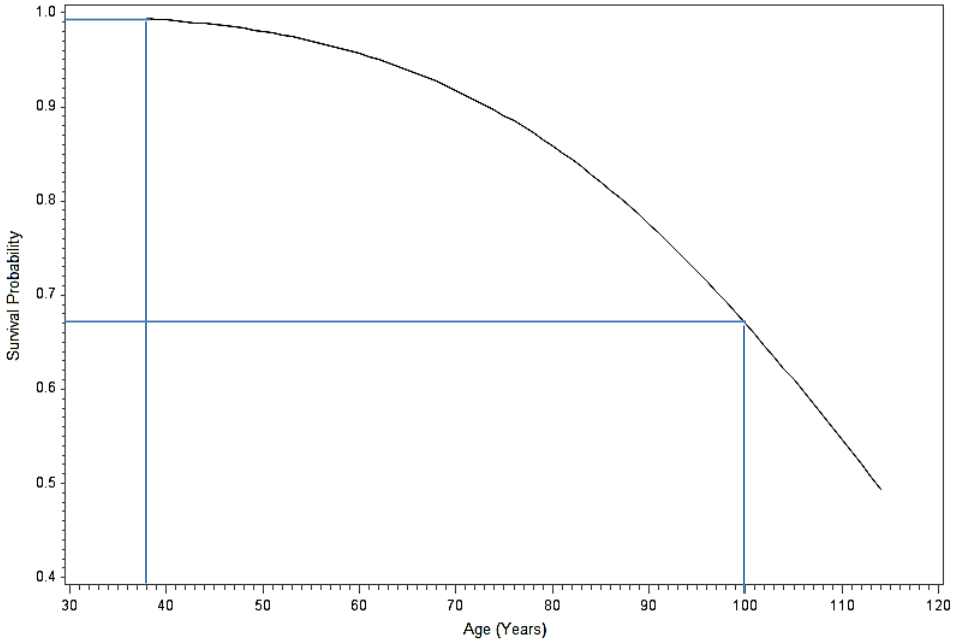


Figure 4.5: Parametric Survival Curve for Water Pipes

Recalling equation 3.2 for a survival curve:

$$S(t) = \exp(-\exp(-\mu\alpha)\, t^{\alpha})$$

where $\mu = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \sigma\varepsilon_i$. Here, $\beta_0, \ldots, \beta_k$ are the estimated parameters, which in this case equals -0.0050, 8.9520, 3.8763, and -0.0007. $x_{i1}, \ldots, x_{ik}$ are the values of k covariates. $\varepsilon_i$is a random error term, and σ is the scale parameter, which in this case equals 0.2314. α is simply $1/\sigma$ or 4.3212.

Therefore, to find the survival probability of a 70 year old water pipe, for example, from the curve in Figure 4.5, the equation would be:

$$S(70) = \exp(-\exp(-4.3212\mu)(70^{4.3212})) \qquad (4.1)$$

It can be seen from Figure 4.5 that the probability of survival of 6" cast iron water pipes is 99% at age 38 and continues to decrease to 67% at age 100.

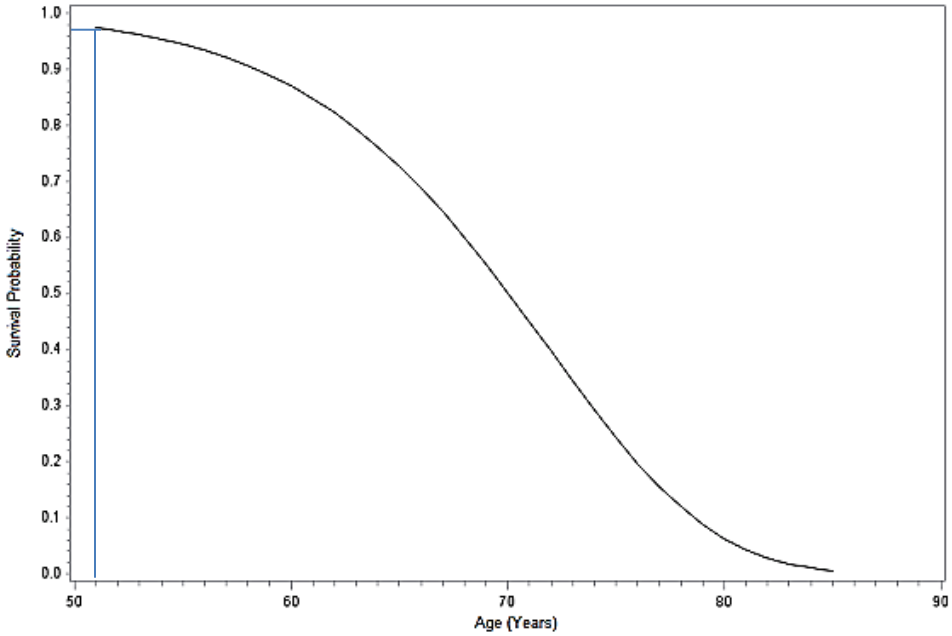


Figure 4.6: Parametric Survival Curve for Bond Hill Water Pipes

Figure 4.6 shows the parametric survival curve for 6" cast iron water pipes from the Bond Hill neighborhood. In this data set, only "C_FACTOR" was the significant variable. It can be seen in Figure 4.6 that the survival probability of this group of pipes is 98% at age 51. It decreases until age 85 where the survival probability becomes 0%. The equation for this curve is:

$$S(t) = \exp(-\exp(-10.4092\mu)(t^{10.4092})) \quad (4.2)$$

The result from selecting significant variables is shown below.

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.6474	0.0795	4.4916	4.8032	3419.13	<.0001
C_FACTOR	1	-0.0076	0.0016	-0.0107	-0.0044	21.79	<.0001
Scale	1	0.0961	0.0111	0.0766	0.1204		
Weibull Shape	1	10.4092	1.2005	8.3033	13.0492		

Figure 4.7: Selection of Significant Variables for Bond Hill Water Pipes

Although we see 2 different curves, a concrete conclusion cannot be made here because the quality of subgroup chosen here can be questioned. This subgroup was only approximately 1% of the main group (48 out of 5,702 records), and any subgroup chosen can coincidently follow the pattern of the 1st curve (5,702 pipes) or be totally contradicting. Nonetheless, these curves obviously show that groups have to be selected carefully because they indeed can have very different deterioration patterns.

4.1.3.2 Non-Parametric Survival Model

For the 5,702 water pipe data, the significant variables were the same as the parametric analysis. They were “C_FACTOR”, “PRESSURE”, “ELEVATION”, and “SHAPE_LENG”. Figure 4.7 shows the non-parametric survival curve for this group of data.

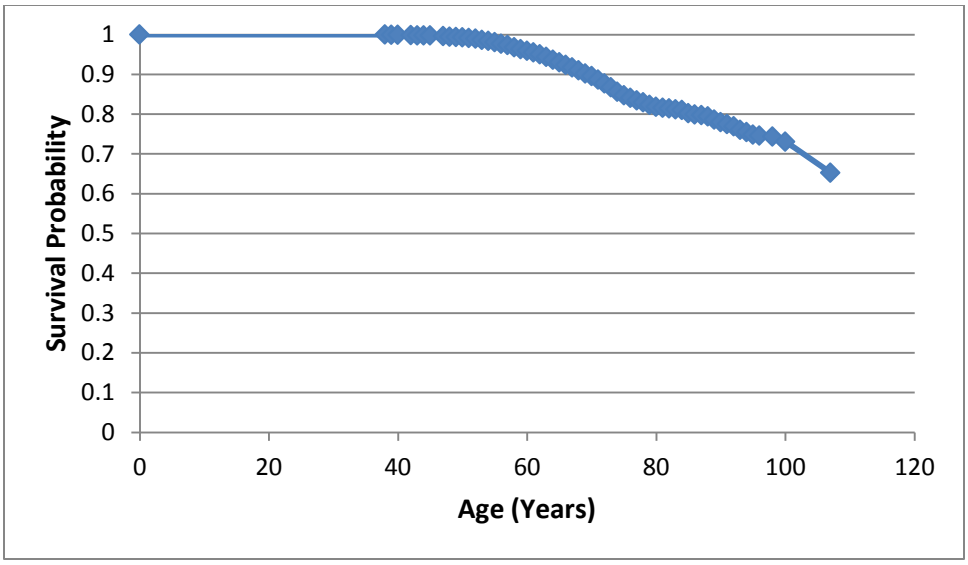


Figure 4.8: Non-Parametric Survival Curve for Water Pipes

It can be seen that the survival probability starts decreasing when the pipes approach approximately age 38. They have a survival probability of 73% at age 100. We see that the survival probability stopped at year 110. This is due to the presence of censored data. Censored data causes the possibility that a curve does not end with zero survival probability.

For the Bond Hill neighborhood data, the significant variables found using the non-parametric analysis were “C_FACTOR” and “SHAPE_LENG”. Figure 4.9 shows the non-parametric survival curve for water pipes in Bond Hill neighborhood.

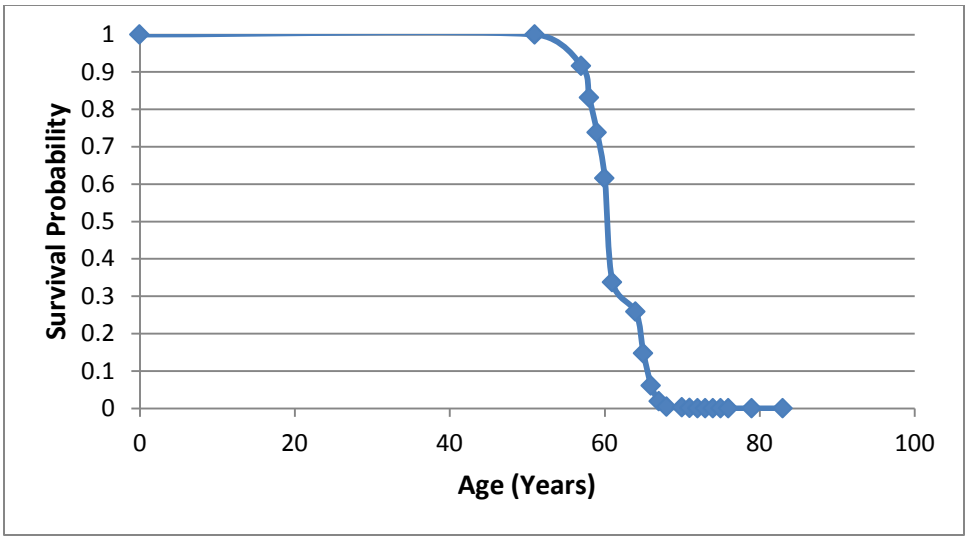


Figure 4.9: Non-Parametric Survival Curve for Bond Hill Water Pipes

It can be seen that the survival probability starts decreasing after the pipes pass 51 years of age. Here, they do approach a 0% survival probability at about age 67.

4.1.4 Model Application

This section discusses applications of the survival curves that were developed for water pipes. It will predict the LoF (or survival probability) of a pipe for a certain number of years, given any age of interest. This helps the asset manager answer management questions presented earlier in chapter 1. However, the CoF score and risk mitigation strategies have to be considered simultaneously to help the asset manager decide whether the risk exposure is within tolerable limits or not.

Figure 4.10 shows the parametric survival curve for water pipes.

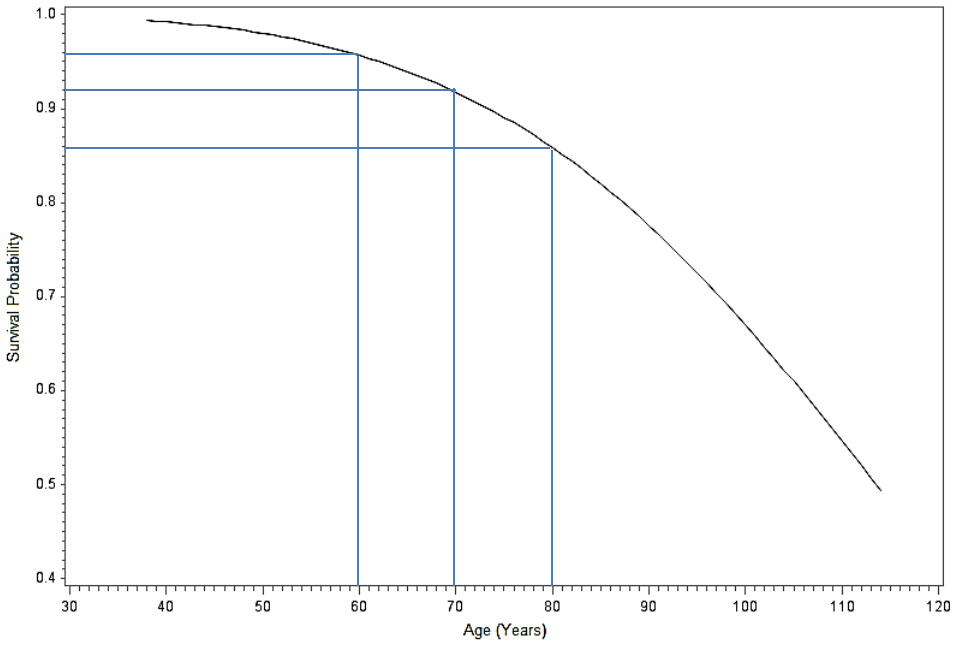


Figure 4.10: Parametric Survival Curve for Water Pipes

The probability of a 60 year old pipe to survive the next 10 years can be calculated as:

$$Probability\ to\ survive\ another\ 10\ years = \frac{Survival\ Probability\ at\ age\ 70}{Survival\ Probability\ at\ age\ 60} = \frac{0.92}{0.96}$$

$$= 95.8\%$$

Similarly, the probability to survive another 20 years at age 60 is:

$$Probability\ to\ survive\ another\ 20\ years = \frac{Survival\ Probability\ at\ age\ 80}{Survival\ Probability\ at\ age\ 60} = \frac{0.86}{0.96}$$

$$= 89.6\%$$

Using this information, the user can prioritize assets available and effectively plan management activities. It can be seen here that the probability of a 60 year old water pipe to survive another 10 years is 95.8% and to survive another 20 years, this probability has only dropped slightly, to 89.6%.

The non-parametric survival curve of water pipes in Figure 4.8 is slightly different from the parametric curve. The graphs are overlapped and shown below.

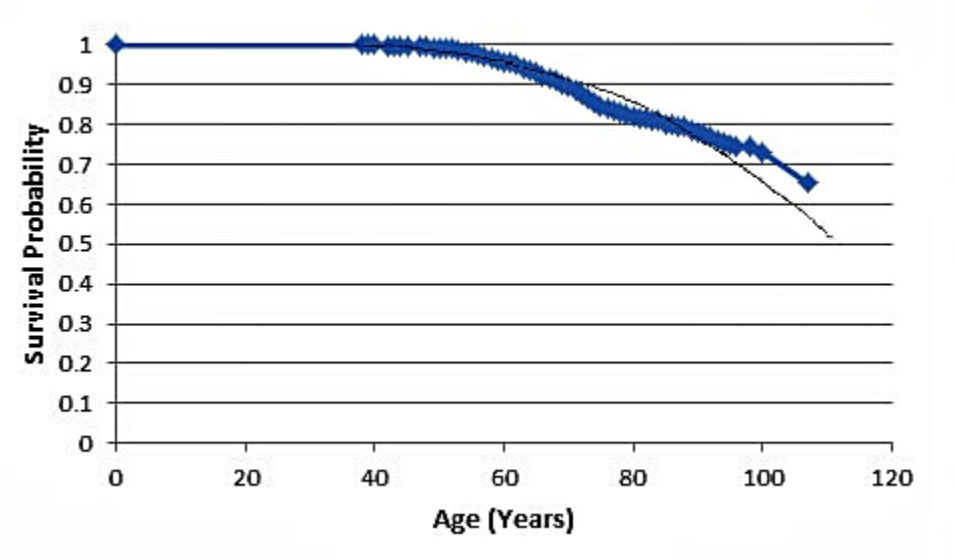


Figure 4.11: Overlapped Survival Curves for Water Pipes

It can be seen that the survival probability shown in the parametric curve decreases at a consistent pattern. However, the non-parametric curve shows the survival probability decreasing in one pattern until age 73 and then it decreases in a different pattern. Nonetheless, they both show that the survival probability starts decreasing at age 38. At age 100, the survival probabilities are 67% and 73%.

Figure 4.12 shows the parametric survival curve for water pipes from Bond Hill neighborhood.

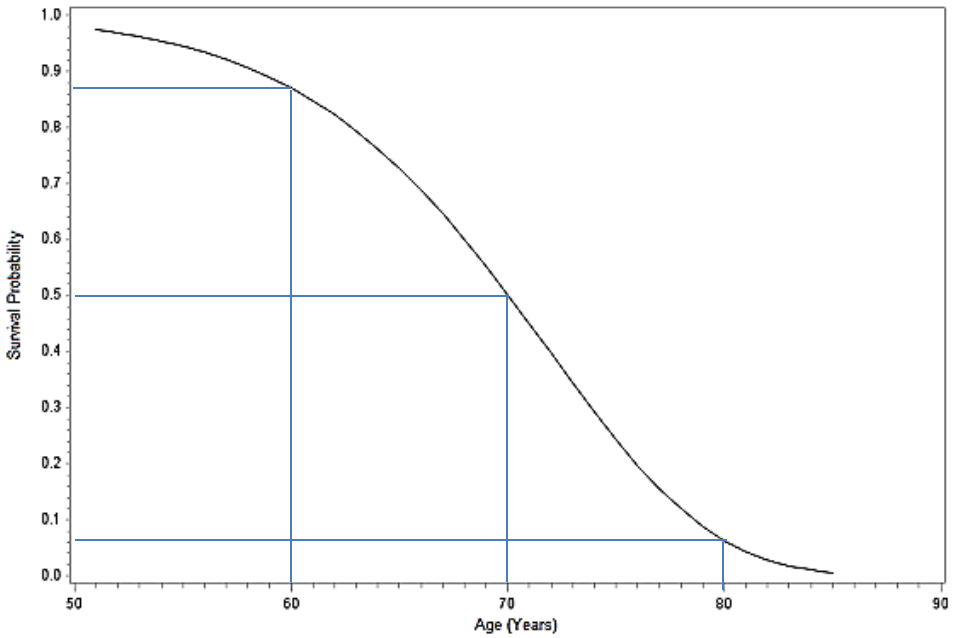


Figure 4.12: Parametric Survival Curve for Bond Hill Water Pipes

The probability of a 60 year old pipe to survive the next 10 years can be calculated as:

$$Probability\ to\ survive\ another\ 10\ years = \frac{Survival\ Probability\ at\ age\ 70}{Survival\ Probability\ at\ age\ 60} = \frac{0.5}{0.87}$$

$$= 57.5\%$$

Similarly, the probability to survive another 20 years at age 60 is:

$$Probability\ to\ survive\ another\ 20\ years = \frac{Survival\ Probability\ at\ age\ 80}{Survival\ Probability\ at\ age\ 60} = \frac{0.06}{0.87}$$

$$= 6.9\%$$

Using this information, the user can prioritize assets available and effectively plan management activities. It can be seen here that the probability of a 60 year old water pipe in Bond Hill neighborhood to survive another 10 years is 57.5% but it only has a 6.9% chance to survive another 20 years.

The non-parametric survival curve of Bond Hill neighborhood water pipes in Figure 4.9 is quite different from the parametric curve in Figure 4.6, and this indicates that the distribution assumed for the parametric model may not be the best. Another possible reason is the nature of the non-parametric model. It is a step curve and due to the small number of data points, the steps can have a very different pattern as when more data points are available. The graphs are overlapped and shown below.

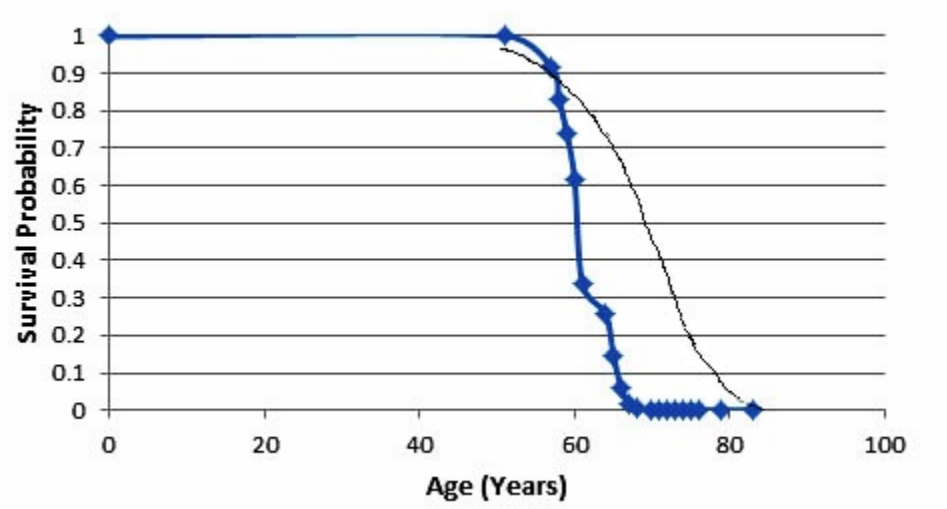


Figure 4.13: Overlapped Survival Curves for Bond Hill Water Pipes

In the non-parametric survival curve, there is extremely little probability (<1%) for a pipe to survive past age 67, whereas the parametric curve shows survivability until age 85. However, both parametric and non-parametric survival curves show that the survival probability starts decreasing at about age 51.

4.2 Sewer Pipe Data

The sewer pipe data was provided by the Metropolitan Sewer District of Greater Cincinnati (MSD). This section presents the analysis results of sewer pipe data. Failure was defined by any maintenance that was done on a pipe. These sewer pipes were only interfered when leaks, collapses, or other unusual symptoms were discovered either from regular inspections or from their users. In this analysis, only the first failures on each pipe were used.

4.2.1 Data Collection, Quality Assurance, and Management

This section describes the data attributes that were collected from MSD data. Table 4.2 shows the data attributes of the sewer data that was collected from MSD.

Table 4.2: Major Data Attributes of Collected Sewer Data

Attribute	**Attribute Type**	**Definition**
Asset Name	Physical	16 digit asset identifier (for example, 28407001-28402014)
COF	Other	Consequence of failure score for the asset
Current Condition	Performance	Current condition rating of the asset at the time of running the analysis
Diameter	Physical	Diameter of the asset in inches
Failure Code	Other	Failure code assigned at time of recorded failure
Failure Date	Operational	Dates of failure
Hierarchy Level 1	Other	Textual description of hierarchy level 1
Hierarchy Level 2	Other	Textual description of hierarchy level 2
Hierarchy Level 3	Other	Textual description of hierarchy level 3
Hierarchy Level 4	Other	Textual description of hierarchy level 4
Hierarchy	Other	Textual description of hierarchy level 5

Level 5		
Install Year	Operational	Year in which the asset was installed
Intervention Condition	Other	Minimum condition score below which the asset cannot be allowed to drop
Length	Physical	Length of pipe segment in feet
Lining Cost	Other	Cost of lining the asset
Material	Physical	Type of pipe segment material (e.g. conc)
Material Class	Other	Class of material that the pipe segment belongs to (e.g. lined, or unlined)
Maximum Life	Other	Maximum life of the pipe segment in years before replacement will occur
Maximum Rehab Count	Other	Maximum possible number of times that the pipe segment can be rehabilitated before being replaced
Max Risk Limit	Other	Maximum risk score that the pipe segment is allowed to reach before appropriate action is triggered
OpCost	Other	Average annual operational costs for the pipe segment
Original Initial Condition	Other	Original condition of the pipe segment without any modifiers
Physical Effective Life	Other	Life of the pipe segment if no rehabilitation was undertaken on it
PmCost	Other	Average annual preventative and predictive maintenance costs for the pipe segment
POF	Other	Probability of failure
Problem code	Other	Code of problem leading to maintenance
PrvRhbCount	Other	Number of times the pipe segment has been rehabilitated in the past
RehabCost	Other	Cost to rehabilitate the pipe segment
RehabDate	Other	Date of last rehabilitation on pipe segment
Renewal Years	Other	Year(s) in which pipe segment was renewed
Replace Action Type	Other	Type of action to be taken during replacement (e.g. open-cut, trenchless, etc.)
Replacement Cost	Other	Cost to replace the pipe segment
RpCost	Other	Cost to repair the pipe segment
Type	Other	Type of asset (e.g. sewer segment, joint, butterfly valve, etc.)

The sewer data had 34 variables. Those that had duplicated information, missing information, irrelevant or were the same for every asset were disregarded. In the end, 5 variables were used in the analysis. They were “Asset Name”, “Diameter”, “Failure Date”, “Install Year”, and “Length”.

4.2.2 Asset Classification

MSD’s master spreadsheet was populated with over 5,000 data points. All the pipes were made of concrete. The diameter distribution is shown in Figure 4.14.

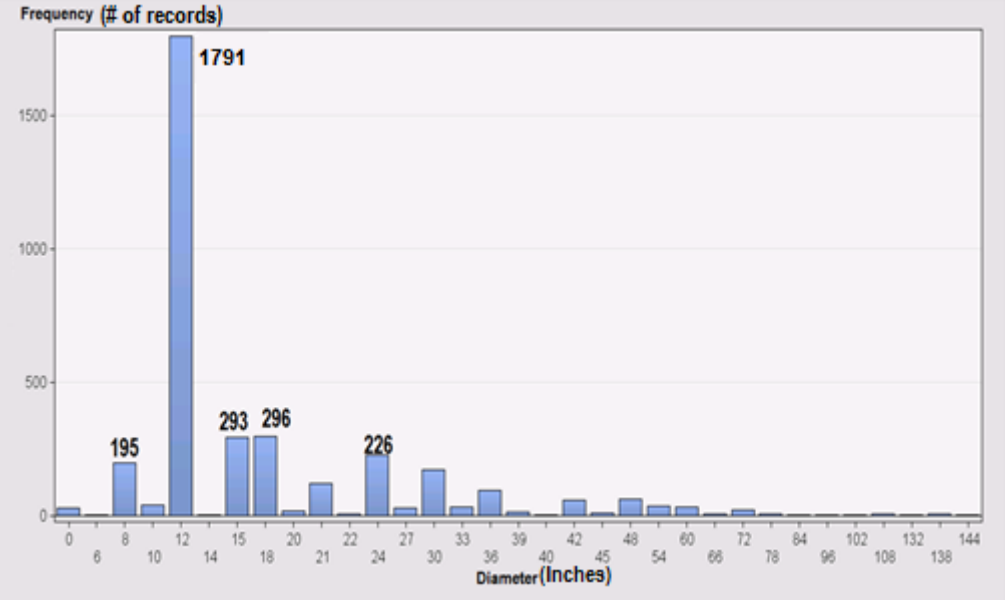


Figure 4.14: Diameter Distribution of Sewer Pipes

The majority of the sewer pipes were 12 inches in diameter. Therefore, this group of pipes was selected for further analysis. There were 1,791 data points.

4.2.3 Survival Model Development

This section shows the results of developing parametric and non-parametric survival models for sewer pipe data.

4.2.3.1 Parametric Survival Model

The 12” concrete sewer pipes had only 1 variable selected, as shown in Figure 4.15. It was tested to be significant. The variable used in this analysis was “LENGTH”.

BEFORE

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.0647	0.0459	4.9749	5.1546	12197.5	<.0001
LENGTH	1	-0.0021	0.0002	-0.0025	-0.0018	132.14	<.0001
Scale	1	0.2770	0.0112	0.2559	0.2999		
Weibull Shape	1	3.6095	0.1459	3.3345	3.9071		

AFTER

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.0647	0.0459	4.9749	5.1546	12197.5	<.0001
LENGTH	1	-0.0021	0.0002	-0.0025	-0.0018	132.14	<.0001
Scale	1	0.2770	0.0112	0.2559	0.2999		
Weibull Shape	1	3.6095	0.1459	3.3345	3.9071		

Figure 4.15: Selection of Significant Variables for Sewer Pipes

Figure 4.16 shows the parametric survival curve for 12” concrete sewer pipes. It can be seen that the probability of survival of these pipes start to decrease at about age 22. At 120 years, the survival probability is approximately 26%.

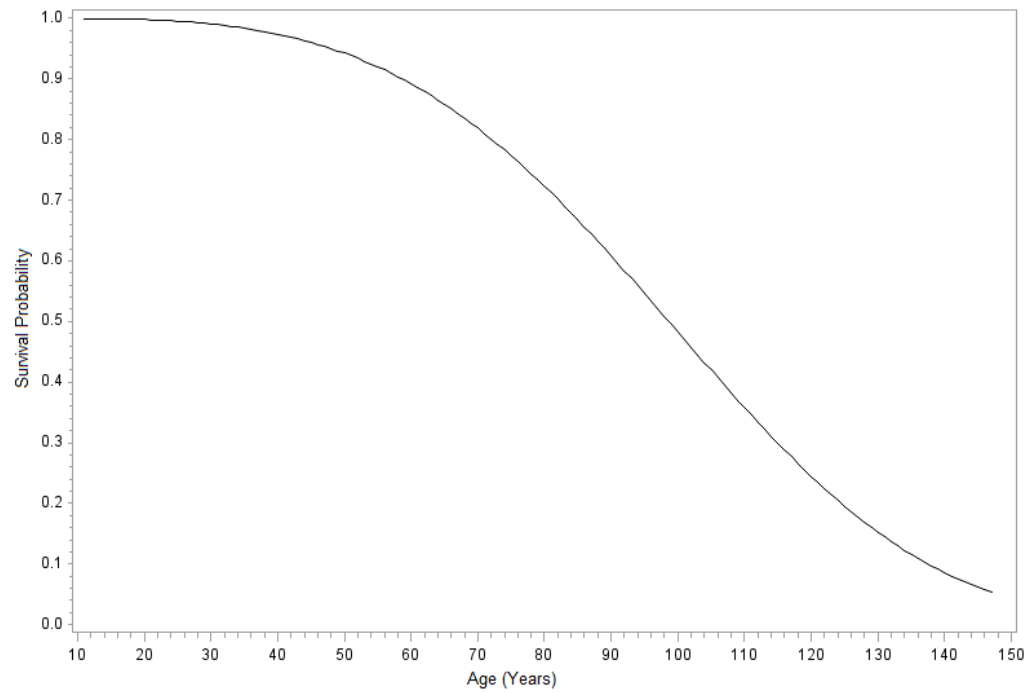


Figure 4.16: Parametric Survival Curve for Sewer Pipes using Variable "Length"

The survival function for the curve in Figure 4.16 would be:

$$S(t) = \exp(-\exp(-3.6095\mu)(t^{3.6095})) \quad (4.3)$$

In an attempt to further study the variable "Length", a new variable "Mod_length" was created whereby pipe lengths were grouped as follows:

Table 4.3: Variable "Mod_length" Groups

Pipe Length (ft.)	**Group**
<100	1
99<Length<200	2
199<Length<300	3
299<Length<400	4

Figure 4.17 shows the new survival curves for these 4 groups of pipes. It can be seen that there are 4 distinct curves for 4 different length categories. The survival probabilities start decreasing at about age 20 and they decrease at different rates. The longer pipes have survival probabilities that decrease at a faster rate compared to shorter pipes.

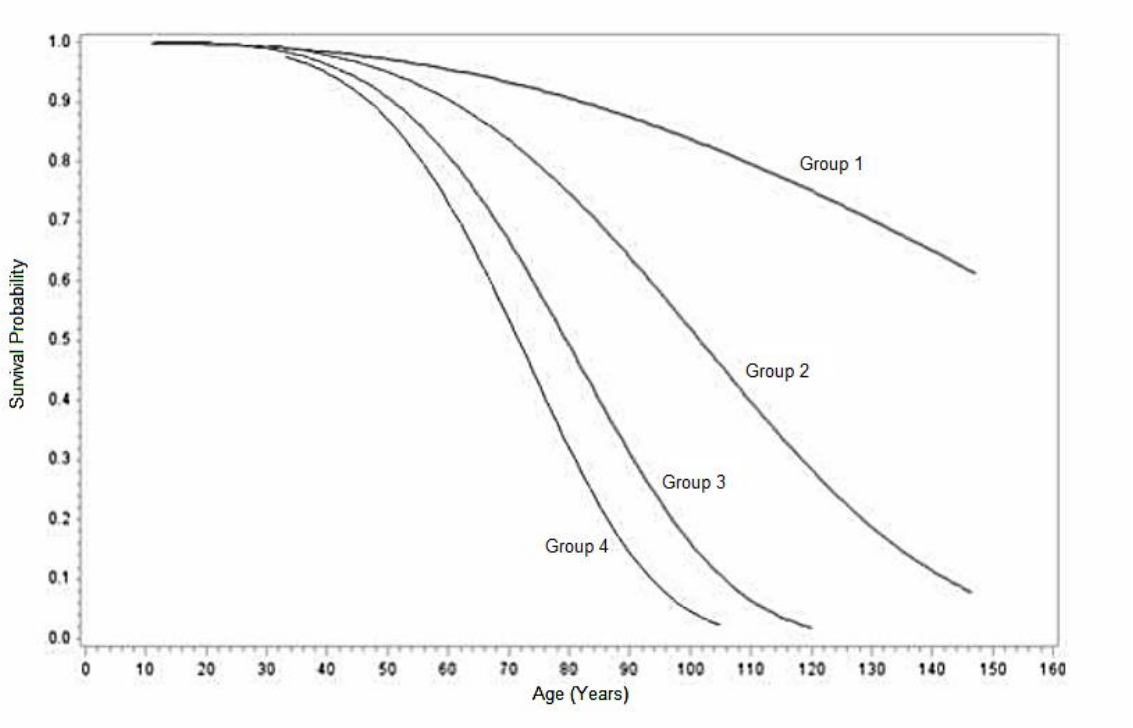


Figure 4.17: Parametric Survival Curves for Sewer Pipes using Variable "Mod_Length"

The equations for these 4 curves are as follows:

$$\text{Group 1: } S(t) = \exp(-\exp(-2.6406\mu)(t^{2.6406})) \quad (4.4)$$

$$\text{Group 2: } S(t) = \exp(-\exp(-3.5961\mu)(t^{3.5961})) \quad (4.5)$$

$$\text{Group 3: } S(t) = \exp(-\exp(-4.2448\mu)(t^{4.2448})) \quad (4.6)$$

$$\text{Group 4: } S(t) = \exp(-\exp(-4.5230\mu)(t^{4.5230})) \quad (4.7)$$

The result tables are shown below.

Group 1	Analysis of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	5.5136	0.2343	5.0544 5.9728	553.87	<.0001
Length	1	-0.0048	0.0028	-0.0102 0.0007	2.97	0.0847
Scale	1	0.3787	0.0510	0.2908 0.4931		
Weibull Shape	1	2.6406	0.3557	2.0280 3.4384		

Group 2	Analysis of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	5.4481	0.1617	5.1313 5.7650	1135.62	<.0001
Length	1	-0.0048	0.0010	-0.0067 -0.0028	23.64	<.0001
Scale	1	0.2781	0.0184	0.2442 0.3166		
Weibull Shape	1	3.5961	0.2380	3.1586 4.0942		

Group 3	Analysis of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	4.6887	0.1973	4.3020 5.0754	564.78	<.0001
Length	1	-0.0009	0.0008	-0.0025 0.0007	1.25	0.2628
Scale	1	0.2356	0.0149	0.2082 0.2666		
Weibull Shape	1	4.2448	0.2677	3.7513 4.8034		

Group 4	Analysis of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	4.1698	0.4248	3.3372 5.0024	96.36	<.0001
Length	1	0.0006	0.0013	-0.0019 0.0030	0.20	0.6571
Scale	1	0.2211	0.0283	0.1720 0.2842		
Weibull Shape	1	4.5230	0.5797	3.5182 5.8146		

Figure 4.18: Result Tables for Sewer Groups

For this study, a hypothesis test was established to test if the survival curves for different pipe lengths were significantly different.

H_0: Survival curves of different pipe lengths were not different.

H_1: Survival curves of different pipe lengths were different

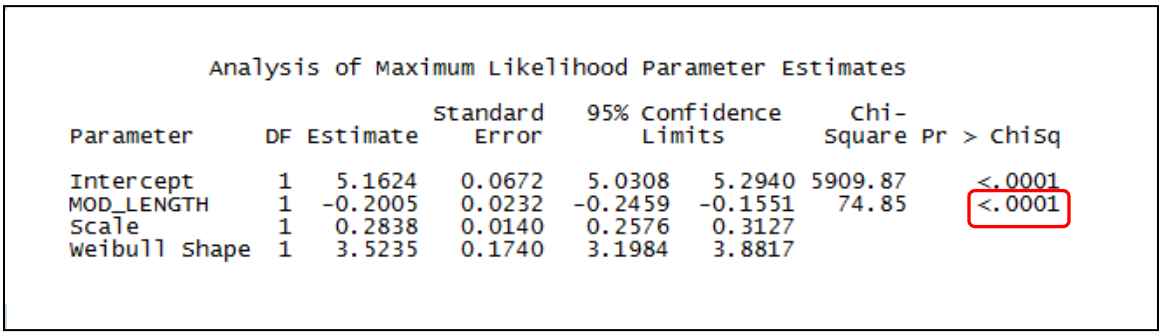
This was done by creating a new variable called “Mod_length” and categorizing the pipe lengths into 5 groups as shown in Table 4.3. A test was then performed on the new variable “Mod_length” to see if the survival curves of different pipe lengths were different. The motivation behind this study was due to the way sewer pipes are being assessed.

The PACP sewer asset condition scoring is the most common scoring system in the United States. There are 3 scoring methods in this system. The Overall Pipe Rating is the total defect score on a pipe. For example, if a pipe has 1 defect score of 5 and 3 defect scores of 4, the Overall Pipe Rating will be 5+4+4+4 = 17. The Quick Rating is a 4 digit number where the first and third digit shows the 2 most severe defect scores. The second and forth digit shows the frequency of these scores. For example, the previous example would have a Quick Rating of 5143. Finally, the Pipe Ratings Index is the average defect score of a pipe. For example, the previous example would have a Pipe Ratings Index of 4.25 (Overall Pipe Rating divided by total number of defects) (Opila, 2011).

The problem with the PACP scoring system is that longer pipes will automatically have poorer scores. This is simply because longer pipes have a higher chance to get more defect scores. More defect scores translate to a worse PACP score. This was the reason “Length” turned out to be a significant variable when analyzing the sewer pipes. Subsequently, special precaution has to be taken when an asset manager makes decisions based on these scores.

As the survival curve for sewer pipes shows, the survival probability of longer pipes decreases at a faster rate compared to shorter pipes. In this analysis, a failed pipe was one that had a defect score of 5. Some utilities would replace the entire pipe segment when in fact only one pipe section had a defect score of 5. Therefore, longer pipes have a higher chance of being classified as a failed pipe and many pipe sections that are still in good condition may be replaced redundantly.

Figure 4.19 shows the analysis results when using variable “Mod_length”.



Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.1624	0.0672	5.0308	5.2940	5909.87	<.0001
MOD_LENGTH	1	-0.2005	0.0232	-0.2459	-0.1551	74.85	<.0001
Scale	1	0.2838	0.0140	0.2576	0.3127		
Weibull Shape	1	3.5235	0.1740	3.1984	3.8817		

Figure 4.19: Results for Testing the Significance of Variable "Mod_length"

The results show that pipe length is indeed sensitive to survival probability ($p<0.0001$). This infers that as the pipe lengths increase from one group to another (100 ft. increments), there is a 18.2% drop in survival time. The calculation is as follows: 100[exp(-0.2005)-1] = -18.2%. In other words, longer pipes have lower probability of survival compared to shorter pipes of the same age. This is a biased judgment based on length. Therefore, there is a need to assess pipes according to their lengths so that such biases would not exist.

To improve the PACP scoring system, defect scores should be assigned according to pipe lengths. The scores should be compared for same lengths of pipes. For example, an asset manager could create a database with pipe segments of similar lengths and assign condition scores. Therefore, no pipe will have a higher chance of getting a poorer score than the other due to being longer. Pipes segments that are longer could be broken down so that there are defect scores for every same length of pipe segments.

4.2.3.2 Non-Parametric Survival Model

Figure 4.20 shows the non-parametric survival curve for sewer pipes when variable “Length” was the significant variable. It can be seen that the survival probability of these pipes starts decreasing at age 33. At age 120, the survival probability reaches 43%.

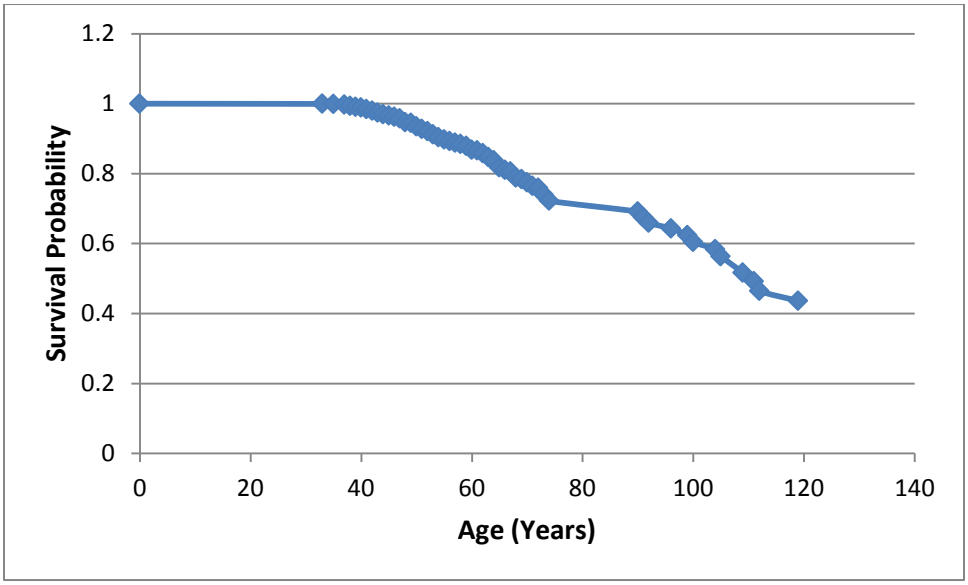


Figure 4.20: Non-Parametric Survival Curve for Sewer Pipes using Variable “Length”

Figure 4.21 shows the survival curve when variable “Mod_length” was used instead of variable “Length”. It can be seen that the curve is identical to Figure 4.20, where variable “Length” was used.

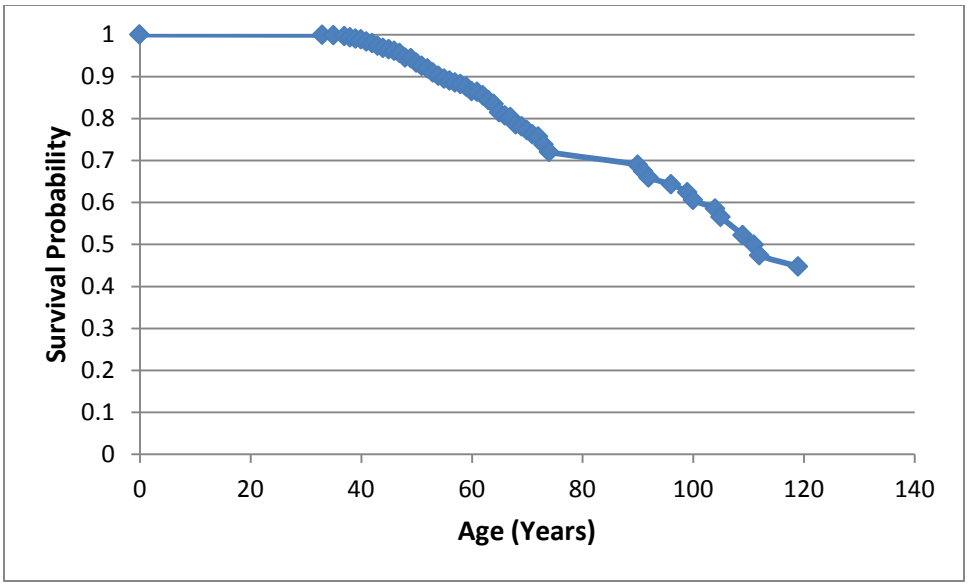


Figure 4.21: Non-Parametric Survival Curve for Sewer Pipes using Variable "Mod_length"

The figure below shows non-parametric curves for the 4 length groups of sewer pipes.

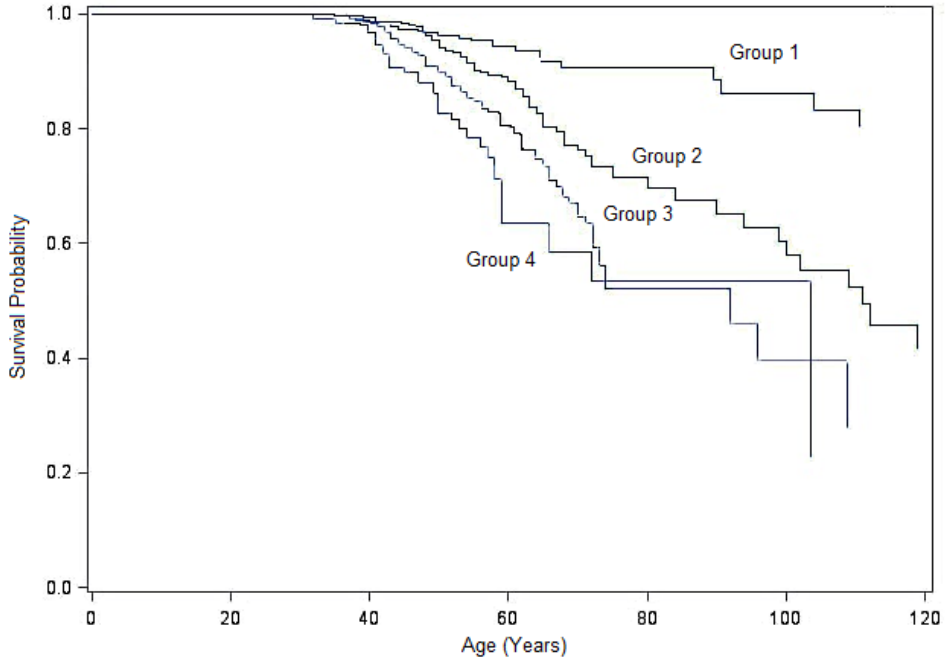


Figure 4.22: Non-Parametric Survival Curves for Sewer Pipes using Variable "Mod_length"

It can clearly be seen that the longer pipes have survival probabilities that drop more rapidly than the shorter ones. We see that the curve for group 4 intersected the curve for group 3, and has a constant survival probability from approximately age 76 to 105. This could be due to limited data for group 4. It has only 131 observations, compared to 398, 672, and 587 observations for groups 1, 2, and 3 respectively.

4.2.4 Model Application

This section discusses applications of the survival curves that were developed for sewer pipes. It will predict the LoF (or survival) of a pipe for a certain number of years, given any age of interest. This helps the asset manager answer management questions presented earlier in chapter

1. However, the CoF score and risk mitigation strategies have to be considered simultaneously to help the asset manager decide whether the risk exposure is within tolerable limits or not.

Figure 4.23 shows the survival curves for 12" concrete sewer pipes using variable "Mod_length". For a pipe that is less than 100 ft. long, its survival probability at age 60 is 95%. At this age, its survival probability for the next 20 years can be calculated as:

$$Probability\ to\ survive\ another\ 20\ years = \frac{Survival\ Probability\ at\ age\ 80}{Survival\ Probability\ at\ age\ 60} = \frac{0.90}{0.95}$$

$$= 94.7\%$$

The survival probability for the next 40 years can be calculated as:

$$Probability\ to\ survive\ another\ 40\ years = \frac{Survival\ Probability\ at\ age\ 100}{Survival\ Probability\ at\ age\ 60} = \frac{0.83}{0.95}$$

$$= 87.4\%$$

Using this information, the user can prioritize assets available and effectively plan management activities. It can be seen that the probability of a 60 year old sewer pipe to survive another 20 years is only 7.3% higher than to survive another 40 years.

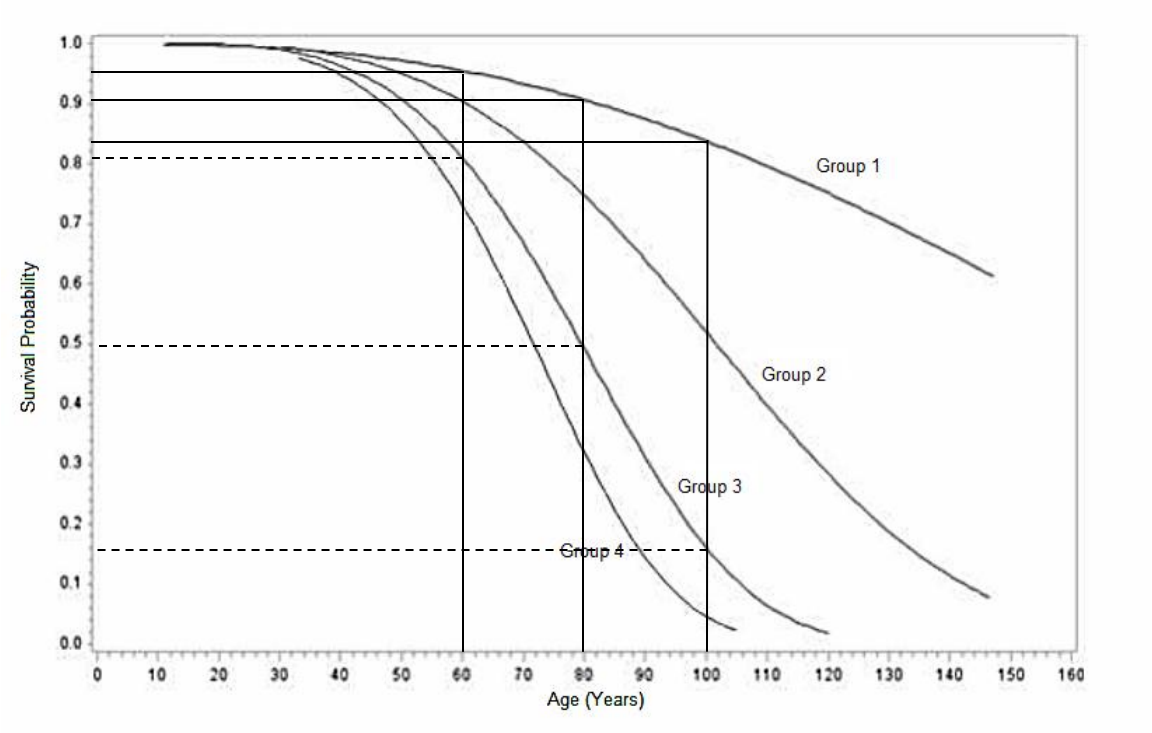


Figure 4.23: Parametric Survival Curves for Sewer Pipes using Variable "Mod_length"

For a pipe that is between 200 ft. and 300 ft., its survival probability at age 60 is 81%. At this age, its survival probability for the next 20 years is calculated as:

$$Probability\ to\ survive\ another\ 20\ years = \frac{Survival\ Probability\ at\ age\ 80}{Survival\ Probability\ at\ age\ 60} = \frac{0.49}{0.81}$$

$$= 60.5\%$$

The survival probability for the next 40 years can be calculated as:

$$Probability\ to\ survive\ another\ 40\ years = \frac{Survival\ Probability\ at\ age\ 100}{Survival\ Probability\ at\ age\ 60} = \frac{0.15}{0.81}$$

$$= 18.5\%$$

Here, the probability of a 60 year old pipe that is less than 100 ft. to survive another 20 years versus 40 years drops by 7.3%. However, the probability of a 60 year old pipe that is 200 ft. to 300 ft. to survive another 20 years versus 40 years drops by 42%.

The non-parametric survival curve for 12" concrete sewer pipes is shown in Figure 4.23. At age 60, its survival probability for the next 20 years can be calculated as:

$$Probability\ to\ survive\ another\ 20\ years = \frac{Survival\ Probability\ at\ age\ 80}{Survival\ Probability\ at\ age\ 60} = \frac{0.70}{0.87}$$

$$= 80.5\%$$

The survival probability for the next 40 years can be calculated as:

$$Probability\ to\ survive\ another\ 40\ years = \frac{Survival\ Probability\ at\ age\ 100}{Survival\ Probability\ at\ age\ 60} = \frac{0.60}{0.87}$$

$$= 68.9\%$$

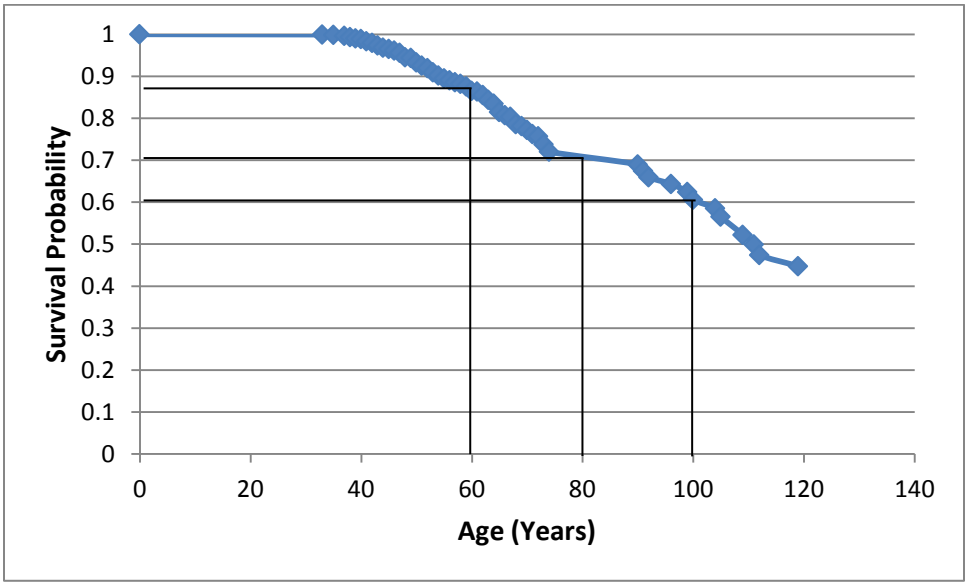


Figure 4.24: Non-Parametric Survival Curve for Sewer Pipes

Here, the probability of a 60 year old sewer pipe to survive another 20 years versus 40 years drops by 11.6%.

4.3 Discussions

This section discusses the results that were obtained from various aspects including data collected and model performances. The distribution assumptions for parametric models will be compared to find the best fitting distribution. Then, the parametric models will also be compared with the non-parametric model.

4.3.1 Data Collected

Both water and sewer data were managed by different organizations. Therefore, they contained different data attributes. Table 4.4 shows what attributes were available from the obtained data versus those that were unavailable.

Table 4.4: Water and Sewer Datasets Comparison

Attribute	**GCWW Data**		**MSD Data**	
	Available	**Unavailable**	**Available**	**Unavailable**
Physical	X		X	
Performance		X	X	
Environmental		X		X
Operational	X		X	

Both datasets met the minimally required attributes to perform the study, which included the physical and operational (installation and failure records) attributes. However, many other physical, environmental, and operational attributes such as soil condition, soil corrosivity, land use type, tree density, proximity to highway, railroads and other structures, etc. that might be obtainable from the engineering department, Geographic Information System (GIS) system and Computerized Maintenance Management System (CMMS) system had not been included in these datasets.

Since this study used survival analysis as a method to determine the LoF and the life expectancy of a pipe, failure data was used rather than condition ratings. Data points that had the physical attributes and installation year but missing failure records were used as censored data points.

In the water data, several additional attributes of interest were available, such as "C_factor", "Pressure", and "Elevation". In the sewer data, attributes such as "Failure Code", "Intervention Condition", "PrvRhbCount", and "Renewal Years" were created for future use. The utility believed that keeping track of these data attributes would help in planning rehabilitation and replacement activities in the future.

The following data points were disregarded in this study:

- Those missing installation year records
- Those without pipe material and diameter information
- Duplicate records

4.3.2 Model Performances and Validation

The table below shows the performances of various distribution assumptions used on water pipe data that were not grouped. The water pipe data had 5,702 data points, and the parametric model was tested using Weibull, Exponential, Lognormal, Gamma, and Loglogistic distributions. The table summarizes the fit statistics using "-2 Log Likelihood", and "AIC". Smaller numbers infer a better fit. It can be seen that for the water data, the gamma distribution fitted the best (-2 log likelihood of 2798.319). It was also compared with the non-parametric model. The non-parametric model did not fit the data well.

Table 4.5: Water Data Fit Statistics

Distribution	**-2 Log Likelihood**	**AIC**
Weibull	3020.001	3032.001
Exponential	4371.847	4381.847
Lognormal	2870.278	2882.278
Gamma	2798.319	2812.319
Loglogistic	2951.451	2963.451
Non-Parametric	18809.991	18817.991

The table below shows the fit statistics for water pipe data that was grouped. Bond Hill data was a subset of data points from the water data set. It contains the records of pipes from the Bond Hill neighborhood. As shown in the table, the models produced fitted much better than the models for the water pipe data that were not grouped. This was partly because Bond Hill data set was smaller. More importantly, the data should fit better because the smaller data set represented a group of pipes that behaved more similarly than in the pool. From the analysis results, we see very different patterns between the plot generated from the pool and the one from Bond Hill neighborhood. This means that the first plot could be generic, and that further classification was necessary. From the Bond Hill data results, it can be seen that the Gamma distribution was also

the best fit among all distribution assumptions (-2 log likelihood of -44.135), and the non-parametric model.

Table 4.6: Bond Hill Water Data Fit Statistics

Distribution	**-2 Log Likelihood**	**AIC**
Weibull	-34.252	-28.252
Exponential	68.744	72.744
Lognormal	-41.858	-35.858
Gamma	-44.135	-36.135
Loglogistic	-41.354	-35.354
Non-Parametric	244.619	246.619

The figures below illustrate a graphical method to test how well an assumed distribution fits the model. It is a cross-check against the fit statistics presented above. Here, the Bond Hill data was tested. The crosses are estimates for the non-parametric model, while the solid straight-sloped line represents the parametric survival model. The dotted lines show the 95% confidence interval limits. Ideally, the non-parametric estimates would match the parametric model. From the figures below, the Exponential model is definitely not a good fit. However, it is difficult to tell which of the other four models is best.

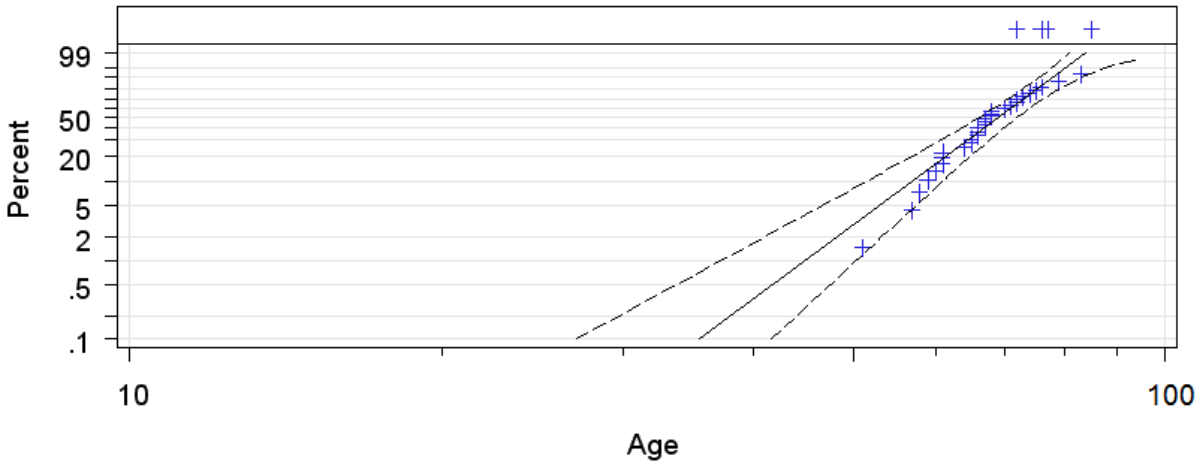


Figure 4.25: Probability Plot for Weibull Distribution

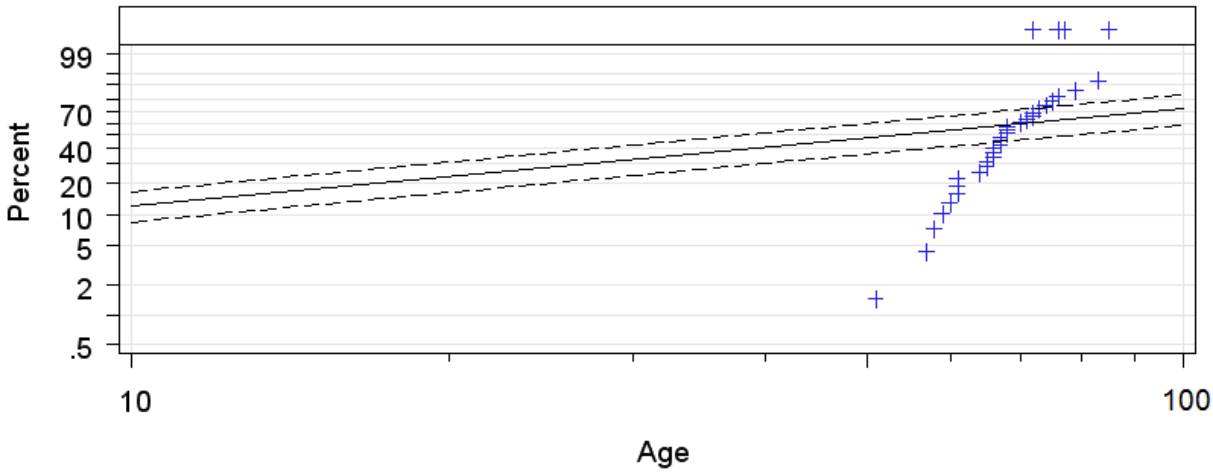


Figure 4.26: Probability Plot for Exponential Distribution

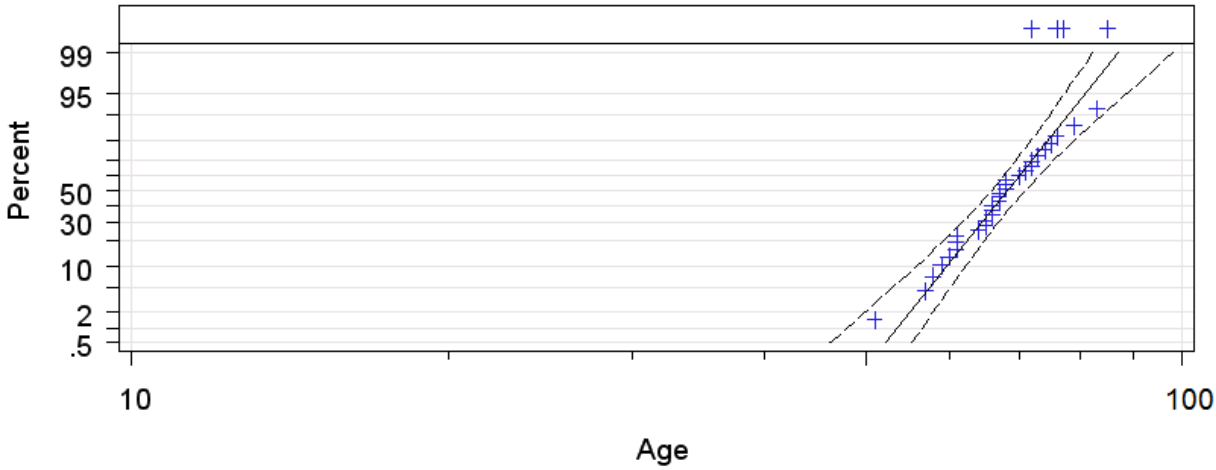


Figure 4.27: Probability Plot for Lognormal Distribution

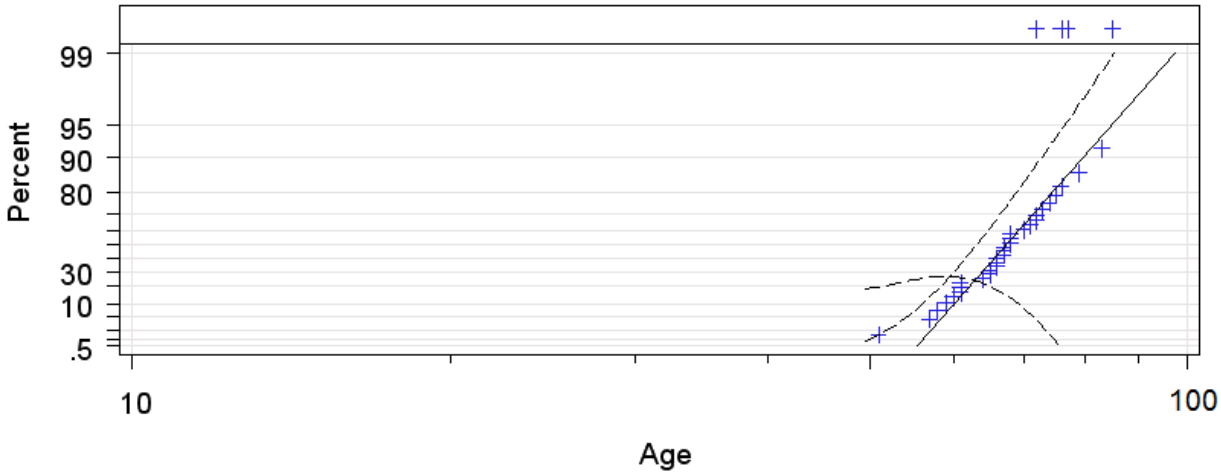


Figure 4.28: Probability Plot for Gamma Distribution

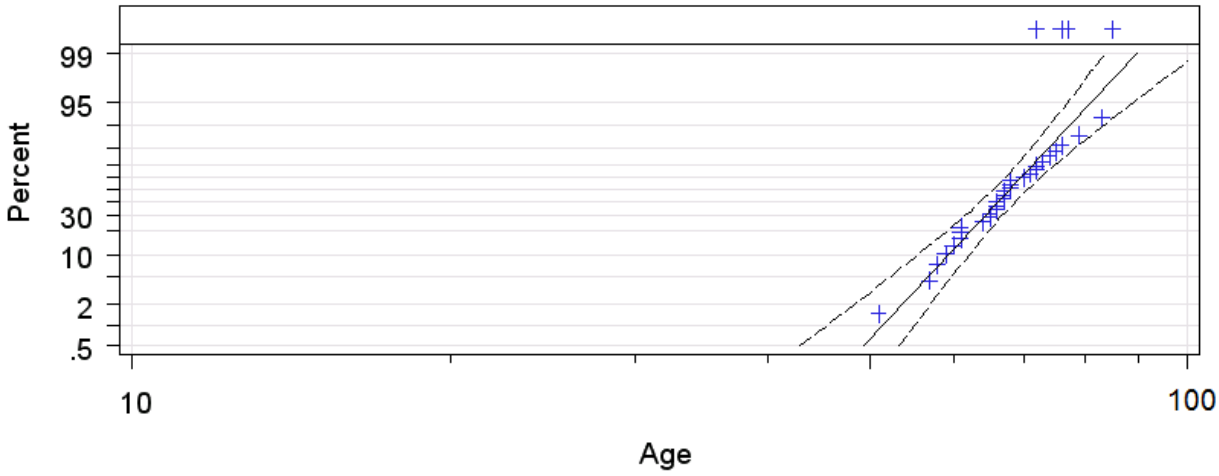


Figure 4.29: Probability Plot for Loglogistic Distribution

The table below shows the fit statistics for sewer data. The sewer data comprised of 1,791 data points. It can be seen from the fit statistics that the gamma distribution produced the best model for these pipes (-2 Log Likelihood of 703.973).

Table 4.7: Sewer Data Fit Statistics

Distribution	**-2 Log Likelihood**	**AIC**
Weibull	761.369	767.369
Exponential	1063.456	1067.456
Lognormal	722.855	728.855
Gamma	703.973	711.973
Loglogistic	743.378	749.378
Non-Parametric	3525.772	3527.772

Gamma distribution could very well produce the best results repeatedly because it is a big family of distributions. In other words, it is much related to the other distributions such as Weibull, Exponential, Lognormal and Loglogistic. However, from these statistics, the Exponential and Non-Parametric distributions can consistently be eliminated.

CHAPTER V

CONSLUSIONS AND RECOMMENDATIONS

There are a number of conclusions that can be drawn from this research. Survival analysis is a great tool because it accounts for censored data, which is a common situation with underground assets. It uses incomplete information to generate results in a way that is consistent and can be trusted. It also accounts for time-dependent data, which is a major advantage over other tools. Here, the parametric models performed better than the non-parametric models, for these underground water and sewer pipes. The fit statistics for parametric models were at least three times better than that of the non-parametric models. Specifically, from the fit statistics, the gamma distribution appeared to be superior for all the data sets. This is probably because it is a big family of distributions. However, the Weibull, Lognormal and Loglogistic distributions seemed to fit well too and may be chosen.

Next, assets need to be grouped. It can be seen that the survival curves for all data sets were different. By creating asset groups, a survival curve that fits better can be created. This can help answer questions such as:

- Which assets, and how likely will they fail this year, or in a particular future year?
- Should current operations and maintenance activities be improved or should an asset be replaced now?
- Should investments be proactive or reactive?

- When and how much should be invested in inspections and condition assessments?
- How can changes in risks be quantified?

However, the size of asset groups needs to be considered because larger groups yield easier management with less survival curves and vice versa.

Thirdly, sewer length is sensitive to survival time due to the way they are assessed. Sewer pipes are scored according to pipe segments, which vary in length. Therefore, longer pipe segments will probably have more defect scores and will have a higher chance of being rated poorer than shorter pipe segments. Results show that for every 100 ft. increment in pipe length, there is an 18.2% drop in survival time.

Recommendations for asset managers include:

- Apply survival analysis to be able to make use of incomplete information. Instead of discarding the entire information on an asset, survival analysis can utilize partial information that is collected and produce reliable results.
- Classify assets strategically using expert judgment or statistical tools. Classifying assets that deteriorate in a similar manner will help identify when and which assets need attention. It will also help answer many asset management questions and improve the effectiveness of the management strategy.
- Assess sewer pipes according to length groups. This will eliminate the bias of longer pipes having lower survival probability.

Recommendations for future researchers in this area include studying the effect of various sample sizes on both parametric and non-parametric models. It is also recommended that the assets be classified statistically, so that the MSGs can be numerically verified. Finally, simpler statistical software may be used so that future researchers will not have to face problems with the software

itself. In this project, a lot of time was spent learning how to use SAS. There were many functions and codes to learn regarding survival analysis, while it was not very easy to find references specifically for this topic. If a freeware such as “R” was used, references might be more easily available.

REFERENCES

[1] Powell, A. E. (2010). "The Infrastructure Roundtables: Seeking Solutions to an American Crisis." *Civil Engineering*, American Society of Civil Engineers.

[2] GHD (2010). "End of Asset Life Reinvestment Decision Making Process Tool: WERF INFR 2010 Research Proposal." GHD.

[3] USEPA (2009). "National Pollutant Discharge Elimination System (NPDES)." <http://cfpub.epa.gov/npdes/index.cfm>. (September 12, 2011).

[4] ASCE (2009). "Report Card for America's Infrastructure: Full Report."Washington, D.C.

[5] Santora, M., and Wilson, R. (2008). "Water Infrastructure in Crisis." *Public Management*, International City/County Management Association, Washington, DC, 17-20.

[6] GHD "Business Risk Exposure Tool."

[7] Marlow, D., Davis, P., Trans, D., Beale, D., and Burn, S. (2009). "Remaining Asset Life: A State of the Art Review." A. Urquhart, ed.

[8] Shamir, U., and Howard, C. D. D. (1979). "An Analytic Approach to Scheduling Pipe Replacement." *Journal of the AWWA*, 248-258.

[9] Herz, R. K. (1996). "Aging Processes and Rehabilitation Needs of Drinking Water Distribution Networks." *Journal of Water, SRT-Aqua*, 45, 221-231.

[10] Duchesne, S., Beardsell, G., Villeneuve, J.-P., Toumbou, B., and Bouchard, K. (2012). "Survival Analysis Model for Sewer Pipe Structural Deterioration: Development and Application." *Computer-Aided Civil and Infrastructure Engineering*Quebec, 1-37.

[11] Syachrani, S. (2010). "Advanced Sewer Asset Management Using Dynamic Deterioration Models." Doctor of Philosophy Dissertation, Oklahoma State University, Stillwater.

[12] Davies, J. P., Clarke, B. A., Whiter, J. T., Cunningham, R. J., and Leidi, A. (2001). "The structural condition of rigid sewer pipes: a statistical investigation." *Urban Water*, 3(4), 277-286.

[13] Ariaratnam, S. T., El-Assaly, A., and Yang, Y. (2001). "Assessment of Infrastructure Inspection Needs Using Logistic Models." *Journal of Infrastructure Systems*, 7(4), 160-165.

[14] Baik, H.-S., Jeong, H. S., and Abraham, D. M. (2006). "Estimating Transition Probabilities in Markov Chain-Based Deterioration Models for Management of Wastewater Systems." *Journal of Water Resources Planning & Management*(January/February), 15-24.

[15] Kulkarni, R. B., Golabi, K., and Chuang, J. (1986). "Analytical techniques for selection of repair-or-replace options for cast-iron gas piping systems--Phase I. Topical report, March 1985-June 1986." Medium: X; Size: Pages: 83.

[16] Shamir, U., and Howard, D. (1979). "An Analytic Approach to Scheduling Pipe Replacement." *Journal of the AWWA*(May 1979), 248-258.
[17] Clark, R. M., Stafford, C. L., and Goodrich, J. A. (1982). "Water Distribution Systems: A Spatial and Cost Evaluation." *Journal of the Water Resources Planning and Management Division*, 108(3), 243-256.
[18] Davis, P., De Silva, D., Marlow, D., Moglia, M., Gould, S., and Burn, S. (2008). "Service Life Prediction and Scheduling Interventions in Asbestos Cement Pipelines." *Journal of Water Supply: Research and Technology - AQUA*, 57(4), 239-252.
[19] Tran, D. H., Ng, A. W. M., and Perera, B. J. C. (2007). "Neural networks deterioration models for serviceability condition of buried stormwater pipes." *Engineering Applications of Artificial Intelligence*, 20(8), 1144-1151.
[20] Kleiner, Y., Sadiq, R., and Rajani, B. (2006). "Modeling the Deterioration of Buried Infrastructure as a Fuzzy Markov Process." *Journal of Water Supply Research Technology: Aqua*, 55(2), 67-80.
[21] Kleiner, Y., Sadiq, R., and Rajani, B. "Sewerage Infrastructure: Fuzzy Techniques to Manage Failures." *Proc., NATO Advanced Research Workshop on Water Reuse: Risk Assessment, Decision-Making and Environmental Security*, 241-252.
[22] Elizabeth Ehret, P. (2011). "Sewer Condition Assessment & Rehabilitation."
[23] NIST (2012). "e-Handbook of Statistical Methods."
[24] Opila, M. C. (2011). "Structural Condition Scoring of Buried Sewer Pipes for Risk-Based Decision Making." Doctor of Philosophy Dissertation, University of Delaware.

APPENDICES

APPENDIX A: LIST OF ACRONYMS

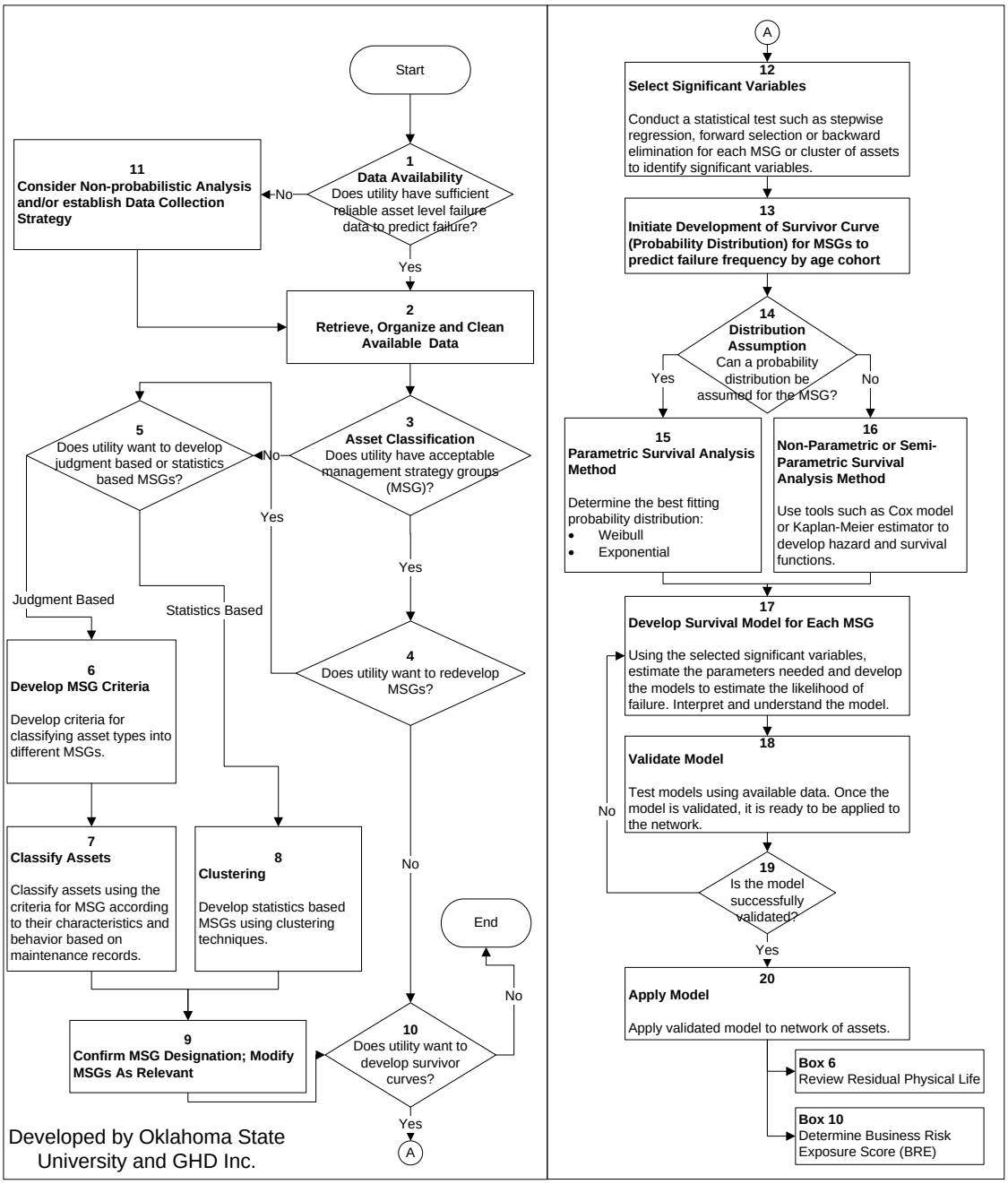
Acronym	Description
ASCE	American Society of Civil Engineers
CoF	Consequence of failure
CWA	Clean Water Act
GCWW	Greater Cincinnati Water Works
GHD	Name of private consultant
LoF	Likelihood of failure
MSD	Metropolitan Sewer District of Greater Cincinnati
MSG	Management Strategy Groups
PACP	Pipeline Assessment Certification Program
ROW	Right of Way
USEPA	United States Environmental Protection Agency

APPENDIX B: REVIEWING LOF SCORE FRAMEWORK

Developing and Applying "Management Strategy Groups" and "Conditional Probabilities" to Determine End of Physical Life

Group Assets by Failure Behavior

Develop Survivor Curves; Project Failure Distribution



VITA

See Hyiik Ting

Candidate for the Degree of

Master of Science

Thesis: PREDICTION OF LIKELIHOOD OF FAILURE OF UNDERGROUND LINEAR ASSETS USING SURVIVAL ANALYSIS

Major Field: Civil Engineering

Biographical:

Education:

Completed the requirements for the Bachelor of Science in Civil Engineering at University at Buffalo, the State University of New York, Buffalo, New York in 2009.

Name: See Hyiik Ting Date of Degree: December, 2012

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: PREDICTION OF LIKELIHOOD OF FAILURE OF UNDERGROUND LINEAR ASSETS USING SURVIVAL ANALYSIS

Pages in Study: 95 Candidate for the Degree of Master of Science

Major Field: Civil Engineering

Scope and Method of Study: Water and sewer pipe assets are challenging to manage because they are buried in the ground. The scope of this paper was to determine the likelihood of failure (LoF) of water and sewer pipes, which was mainly their physical failure. The pipes that possibly deteriorated similarly were grouped either by judgmental or statistical methods. Survival analysis was used for this study due to its ability to include censored data, which was common for underground assets that do not get inspected very often and are difficult to keep track of. Parametric and non-parametric models were developed.

Findings and Conclusions: The parametric model is better in predicting the LoF for underground assets. For all the data sets tested, the gamma distribution fitted the best. It is essential to manage assets in groups. By grouping similarly behaved pipes together, survival curves can be developed to predict their LoF effectively. Smaller asset groups would lead to many survival curves and it could be difficult to manage. However, accurate survival curves may be difficult to generate for large groups of assets. A compromise has to be made for the right group sizes so that an asset manager can effectively manage his assets. While studying the sewer pipes, it was found that for every 100 ft. increments in pipe lengths, there was an 18.2% drop in survival probability. This problem could be solved if pipe conditions were assessed in similar length groups. The survival curve is different for every group of assets.

ADVISER'S APPROVAL: Dr. Hyung Seok Jeong